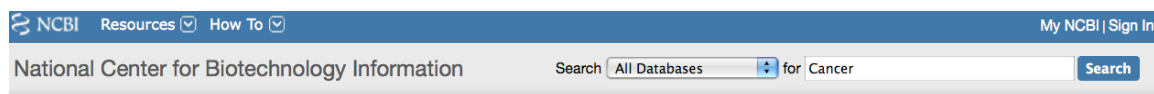


NCBI Exercises Set 1

Entrez Exercises

Global Query: Controlled Vocabularies and Limits

Type the word “cancer” in the search box on the NCBI homepage and run the search.

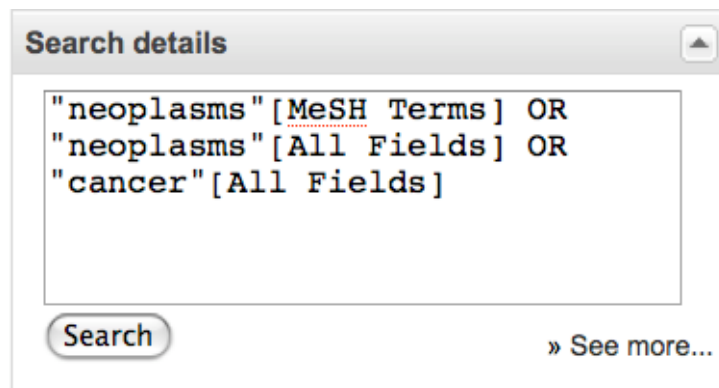


This query returns results in all of the Entrez databases. However the query is interpreted differently in different databases.

PubMed

Retrieve the result for the PubMed database. Look at the “Search details” in the right-hand column to see how the query was interpreted in this database.

Notice that the term cancer was translated to the Medical Subject Heading (MeSH) term “neoplasms” (“neoplasms”[MeSH Terms]).



MeSH is a controlled vocabulary that is used to index all articles in PubMed. In the details box, edit the query to remove the portion that searched for cancer as a text word and run the search. Notice that the number of articles retrieved has changed. These will be a more relevant set of results.

You can force the PubMed engine to only search the MeSH vocabulary by editing the query to only search MeSH Terms.

Edit the query in the “Search details” so that only the “neoplasms”[MeSH Terms] query remains and click the “Search button”.

Now run the search with the limit in place and check the “Details” tab to verify that only the MeSH term translation was used.

Nucleotide

The Nucleotide database in the Global query is now three separate databases. The two large bulk sequence divisions, the expressed sequence tags (EST) and the genome survey sequences (GSS) are in their own separate Entrez databases. The remaining sequence records are in the database.

Use the Web browser’s back button to return to the Global query page. Retrieve the results for the Nucleotide database.

Limits Preview/Index History Clipboard Details

Found 4613129 nucleotide sequences. Nucleotide [779932] EST [3522352] GSS [310845]

Display Summary Show 20 Sort By Send to

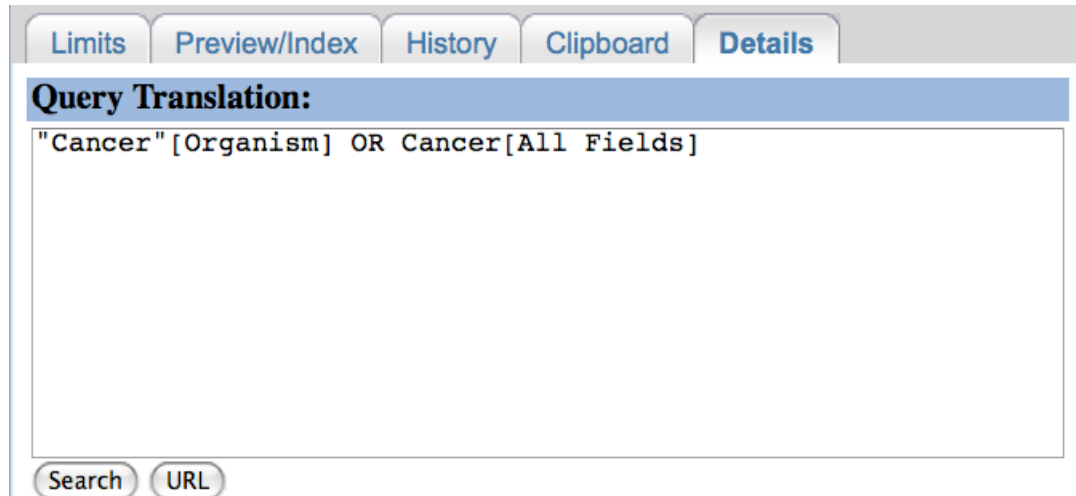
All: 779932 Bacteria: 15756 RefSeq: 47538 mRNA: 229598

Items 1 - 20 of 779932 Page 1 of 38997 Next

☐ [Homo sapiens ADP-ribosylation factor-like tumor suppressor protein 1 \(ARLTS1\) mRNA, complete cds](#)

1. 3,760 bp linear mRNA
AF441378.1 GI:30060409

Click the “Details” tab to see how the query was interpreted for this molecular database.



The screenshot shows a web interface with five tabs: 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. The 'Details' tab is selected and highlighted in blue. Below the tabs is a section titled 'Query Translation:' in a blue header. Inside this section, a text box contains the query: `"Cancer"[Organism] OR Cancer[All Fields]`. Below the text box are two buttons: 'Search' and 'URL'.

In this database, the term cancer was translated into the crustacean genus name *Cancer* ("Cancer"[Organism]). The organism field stores NCBI's taxonomic classification for the source organism for the record. This is the most important controlled vocabulary for the bio-molecular Entrez databases. In this case, this translation has an unintended consequence of retrieving unrelated records: those from the crustacean genus *Cancer*, and those containing the term cancer most often in the context of the disease.

In the details box, edit the query to remove the portion that searched for cancer in all fields so that you are just performing a search with "Cancer"[Organism] and run the search.

This retrieves all of the nucleotide sequences for the genus *Cancer*. As you did with PubMed and the MeSH terms, you can use the "Limits" tab in the bio-molecular databases to restrict your search to the Organism field and obtain only the records from the crab genus.

Taxonomy

Go back to the global query results or run the search again for cancer on the NCBI homepage and retrieve the single result for the taxonomy database and click on the linked name.

This takes you into the taxonomy browser and allows you to see all entries for the genus *Cancer*. You can check the boxes at the top to see the number of records from this genus in the various bio-molecular databases. (You must click the "Display" button to see the numbers.) These numbers are hyperlinks that will retrieve the records from the databases. The taxonomy database and browser are very useful as a global query for organism names in the bio-molecular databases.

◦ [Cancer](#) [97](#) [93](#) [12](#) [80](#) [LinkOut](#) Click on organism name to get more information.

- [Cancer antennarius](#) (Pacific rock crab) [4](#) [4](#) [2](#) [4](#) [LinkOut](#)
- [Cancer borealis](#) (Jonah crab) [11](#) [12](#) [44](#) [LinkOut](#)
- [Cancer branneri](#) (furrowed rock crab) [1](#) [1](#) [1](#) [LinkOut](#)
- [Cancer irroratus](#) (Atlantic rock crab) [18](#) [16](#) [3](#) [7](#) [LinkOut](#)
- [Cancer japonica](#) [2](#) [2](#) [2](#)
- [Cancer jordani](#) [2](#) [2](#) [2](#) [LinkOut](#)
- [Cancer oregonensis](#) (pygmy rock crab) [3](#) [3](#) [1](#) [1](#) [LinkOut](#)
- [Cancer pagurus](#) (edible crab) [38](#) [37](#) [4](#) [31](#) [LinkOut](#)
- [Cancer plebejus](#) [3](#) [3](#) [1](#)
- [Cancer porteri](#) [1](#) [1](#) [1](#)
- [Cancer productus](#) (red rock crab) [14](#) [12](#) [4](#) [15](#) [LinkOut](#)

Nucleotide: Zebrafish prolactin

Zebrafish nucleotide sequences

Perform a search to retrieve all zebrafish sequences in the Nucleotide database. Use the “Limits” tab to select the “Organism” field to force the translation to an organism search.

Limits: the Properties field

You can now use the “Limits” to eliminate certain types of sequences from your results.

Click on “Limits” and use the checkboxes to remove the high throughput genomic (HTG or Working Draft) sequences from your results. Check the box next to “exclude working draft” and run the search.

NCBI Nucleotide

Search Nucleotide for zebrafish Go Clear

My NCBI Sign In Register

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

About Entrez

Entrez Nucleotide Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

My NCBI

Related resources

BLAST

Reference sequence project

Search for Genes

Submit to GenBank

Search for full length cDNAs

Limits Preview/Index History Clipboard Details

Limited to:

Organism: Any

Exclude:

☐ STSs ☒ working draft ☐ TPA ☐ patents

Molecule: Any

Gene Location: Any

Segmented Sequences: Any

Only from: Any

Published in the last: Any Date

Modified in the last: Any Date

Click on the “Details” tab to see how Entrez managed this query.

Notice the term “NOT gbdiv_htg[PROP]”. PROP is the abbreviation for the Properties field

Limits Preview/Index History Clipboard Details

Field: Organism Limits: exclude working draft

Query Translation:

"Danio rerio"[Organism] NOT gbdiv_htg[PROP])

Search URL

Result:

160287

Translations:

zebrafish "Danio rerio"[Organism]

Database:

Nucleotide

The Properties field terms are a controlled vocabulary for classifying sequence records. These terms are somewhat cryptic, but they are very helpful. Three useful types are the `biomol`, `gbdiv` and `srcdb` sets. The `biomol` terms classify records based on the type and origin of the molecule, for example `biomol mrna` or `biomol genomic`. The `gbdiv` sets of terms index records by the GenBank division code; `gbdiv est`, `gbdiv pri`, `gbdiv htg` and so on. The `srcdb` terms classify records based upon their database of origin. For nucleotide records these could be GenBank, EMBL, DDBJ, RefSeq or PDB (`srcdb genbank`, `srcdb embl`, `srcdb ddbj`, `srcdb refseq`). Many of the available filters on the “Limits” tab are managed through the Properties field terms.

Preview/Index: adding terms to query

Return to the CoreNucleotide search results. Go into “Limits” again and use the “Molecule” drop-down menu to select mRNA and run the search.

The results now contain all non-EST zebrafish mRNA sequences from the primary databases and the RefSeq database.

Click on the “Preview/Index” tab.

At the bottom of the “Preview/Index” page, is a search box with a drop-down menu that allows you to add terms to your search and restrict to certain fields if you like.

Now, type “prolactin” in the search box.

Although the vocabulary used is not strictly controlled, the name of a gene or gene product is generally in the title of a record. The title is displayed in the “Summary” view in Entrez and is identical to the DEFINITION line in a record in GenBank format. Select “Title” from the drop-down menu to the left of the search box and click the “Index” button. This checks the index for the “Title” field for records having “prolactin” in their titles. A list containing term prolactin and its expansion is now displayed with the number of records for each term.

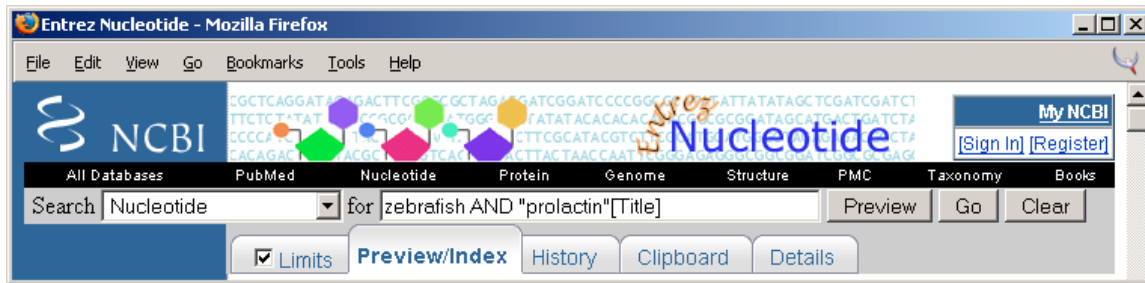
Title Preview Index

Click to add a term to the query box

- prolactin (1174)
- prolactin 1 (3)
- prolactin 2 (20)
- prolactin a (2)
- prolactin and (3)
- prolactin b (2)
- prolactin c (1)
- prolactin f (1)
- prolactin family (115)
- prolactin gene (203)

Up Down

Select “prolactin” from the list and add it to the search by clicking the “AND” button. Then run the search.



The results contain records from GenBank / EMBL / DDBJ and NCBI's RefSeq database. The RefSeq records are easily identified by their characteristic style of accession numbers. Retrieve the RefSeq record for the zebrafish prolactin mRNA (NM_181437). This RefSeq contains sequence data derived from a traditional GenBank record, but also has additional annotations and cross references added by the NCBI RefSeq staff. Unlike primary database records, this RefSeq record will be updated and maintained as the state of knowledge about the biology of this gene and organism advances. The results also contain records for zebrafish prolactin receptors as well as an additional prolactin 2 transcript. Zebrafish are tetraploid and often retain both gene copies from the original duplication as is apparently the case here for the two prolactins and the prolactin receptors. The search also retrieves a gene model RefSeq for the prolactin receptor-like transcript (XM_685247). This sequence has been predicted from analysis of the assembled zebrafish genome using the NCBI gene prediction program called Gnomon.

Finding the genomic BAC clone sequence

Retrieve the record for NM_181437 by clicking on the linked title. Click on the “Links” menu in the upper right of the record (NM_181437). The expanded “Links” menu is also available at the bottom of the right-hand “Discovery” column on the sequence record.

A number of links are displayed. You can link directly to the assembled and annotated whole genome shotgun assembly of the zebrafish genome in the Map Viewer. There are also a growing number of finished BAC clone sequences from the zebrafish genome project that are available in Entrez. You can use the “Related Sequences” feature of Entrez to find a BAC clone that contains

the exons of this gene.

Format: [GenBank](#) [FASTA](#) [Graphics](#) [More Formats](#) ▼

[Download](#) ▼

[Save](#) ▼

[Links](#) ▼

NCBI Reference Sequence: NM_181437.3

Danio rerio prolactin (prl), mRNA

[Comment](#) [Features](#) [Sequence](#)

LOCUS NM_181437 1425 bp mRNA linear VRT
26-JUL-2009
DEFINITION Danio rerio prolactin (prl), mRNA.
ACCESSION NM_181437 XM_001331408
VERSION NM_181437.3 GI:127138930
KEYWORDS .
SOURCE Danio rerio (zebrafish)
ORGANISM [Danio rerio](#)
Euteleostomi;
Actinopterygii; Neopterygii; Teleostei; Ostariophysi;
Cypriniformes; Cyprinidae; Danio.
REFERENCE 1 (bases 1 to 1425)
AUTHORS Nguyen,N. and Zhu,Y.
TITLE Prolactin functions as a survival factor during zebrafish
embryogenesis
JOURNAL Comp. Biochem. Physiol., Part A Mol. Integr. Physiol. 153
(1), 88-93 (2009)
PUBMED [19032987](#)
REMARK GeneRIF: Prolactin acts as a survival factor during zebrafish
embryogenesis.
REFERENCE 2 (bases 1 to 1425)
AUTHORS Dutta,S., Dietrich,J.E., Westerfield,M. and Varga,Z.M.
TITLE Notch signaling regulates endocrine cell specification in the
zebrafish anterior pituitary

[Change Region Settings](#)

[Customize View](#)

Sequence Analysis

► [BLAST Sequence](#)

► [Pick Primers](#)

Articles about the

► [Prolactin function in zebrafish](#)

► [Notch signaling regulates endocrine cell specification in the zebrafish anterior pituitary](#)

► [In vivo time-lapse imaging of zebrafish embryos](#)

[RefSeq Protein](#)

See the reference protein for prolactin (NP_852111.1)

[More about the prl gene](#)

Also Known As: zgc:110500

Full text in PMC

Gene

Genome

Master

Protein

PubMed

PubMed (RefSeq)

PubMed (Weighted)

Taxonomy

mRNA Genome

Related Sequences

Map Viewer

GEO Profiles

UniGene

LinkOut

All links from this record

- [Full text in PMC](#)
- [Gene](#)
- [Genome](#)
- [Master](#)
- [Protein](#)
- [PubMed](#)
- [PubMed \(RefSeq\)](#)
- [PubMed \(Weighted\)](#)
- [Taxonomy](#)
- [mRNA Genome Project](#)
- [Related Sequences](#)
- [Map Viewer](#)
- [GEO Profiles](#)
- [UniGene](#)
- [LinkOut](#)

Follow the “Related Sequences” link.

This provides a list of nucleotide sequences that are related by BLAST similarity. Similarity scores are precomputed between all sequences in the database. The related sequences list is ranked in

order of decreasing BLAST score. For the nucleotide database, the significance threshold is very stringent, so that it is unusual to see nucleotide sequences from other species in the list. Therefore, the nucleotide related sequences link is often a useful as a way of collecting all sequences for a particular gene and its products from one species. Often you can't easily collect all of them using a text search because of inconsistencies or errors in the annotation.

You should find the sequence from BAC clone DKEY-16P21, accession BX511021, in the list of related sequences. Retrieve this record through the linked identifier.

This is a typical finished BAC clone from a genome project. Notice that this is the ninth version of this record. In previous versions, this was a draft sequence in the high throughput genomic (HTG) GenBank division. You can see all versions of the record in the revision history available through the "More Formats" link.

Format: [GenBank](#) [FASTA](#) [Graphics](#) [More Formats](#)▼
Download▼ Save▼ Links▼

GenBank: BX511021.9

GenBank(Full)
ASN.1
Revision History

Zebrafish DNA sequence from clone DKEY-16P21 in linkage group 3, complete sequence.

Change Region Shown▼
Customize View▼

[Comment](#) [Features](#) [Sequence](#)

LOCUS BX511021 235632 bp DNA linear VRT
03-APR-2004

DEFINITION Zebrafish DNA sequence from clone DKEY-16P21 in linkage group 3, complete sequence.

ACCESSION BX511021

VERSION BX511021.9 GI:46200452

KEYWORDS HTG.

SOURCE Danio rerio (zebrafish)

ORGANISM [Danio rerio](#)
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Ostariophysi; Cypriniformes; Cyprinidae; Danio.

REFERENCE 1 (bases 1 to 235632)

AUTHORS Hunter,G.

TITLE Direct Submission

JOURNAL Submitted (03-APR-2004) Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, UK. E-mail enquiries: zfish-help@sanger.ac.uk Clone requests: clonerequest@sanger.ac.uk

COMMENT On Apr 5, 2004 this sequence version replaced gi:[44844493](#).

Sequence Analysis Tools
▶ BLAST Sequence
▶ Pick Primers

Recent activity▼

All links from this record

- ▶ Assembly
- ▶ Gene
- ▶ Taxonomy
- ▶ Related Sequences
- ▶ Map Viewer
- ▶ SNP
- ▶ UniSTS
- ▶ LinkOut

The revision history for this record shows all the forms it has taken in the Entrez system.

Sequence Revision History

Find (Accessions, GI numbers or Fasta style Seq(ds))

About Entrez difference between I and II as

Revision history for [BX511021](#)

GI	Version	Update Date	Status	I	II
46200452	9	Oct 20 2006 2:13 PM	Live	<input checked="" type="radio"/>	<input type="radio"/>
46200452	9	Apr 5 2004 12:43 PM	Dead	<input type="radio"/>	<input checked="" type="radio"/>
44844493	8	Mar 1 2004 11:12 PM	Dead	<input type="radio"/>	<input checked="" type="radio"/>
42820886	7	Feb 25 2004 11:07 PM	Dead	<input type="radio"/>	<input checked="" type="radio"/>
42733255	6	Feb 22 2004 11:18 PM	Dead	<input type="radio"/>	<input checked="" type="radio"/>
42733255	6	Feb 20 2004 11:16 PM	Dead	<input type="radio"/>	<input checked="" type="radio"/>
42592613	5	Feb 18 2004 11:07 PM	Dead	<input type="radio"/>	<input checked="" type="radio"/>
42592613	5	Feb 17 2004 11:05 PM	Dead	<input type="radio"/>	<input checked="" type="radio"/>
42538790	4	Feb 11 2004 11:13 PM	Dead	<input type="radio"/>	<input checked="" type="radio"/>
32959715	3	Jul 17 2003 11:57 PM	Dead	<input type="radio"/>	<input checked="" type="radio"/>
31071334	2	Jul 1 2003 11:29 PM	Dead	<input type="radio"/>	<input checked="" type="radio"/>
31071334	2	May 23 2003 11:13 PM	Dead	<input type="radio"/>	<input checked="" type="radio"/>
30910895	1	May 19 2003 11:07 PM	Dead	<input type="radio"/>	<input checked="" type="radio"/>

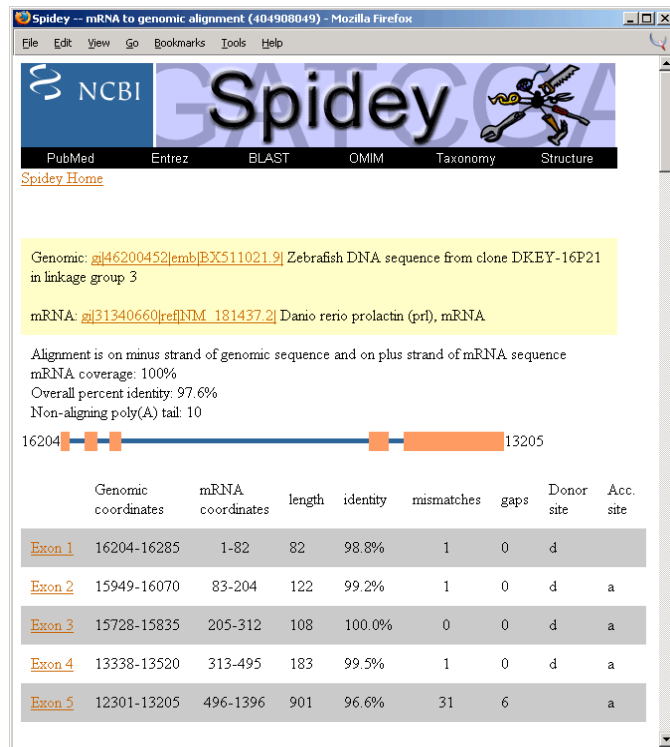
Accession [BX511021](#) was first seen at NCBI on May 19 2003 11:07 PM

The revision history also shows the gi number and accession.version number changes. These identifiers change together and only when the sequence itself changes. Other non –sequence changes can be made that do not affect these identifiers but are reflected in changes in the “Update Date”.

The current version of BX511021 is no longer a draft sequence but is in the traditional vertebrate (VRT) division of GenBank. In contrast to typical traditional GenBank records, BX511021 has almost no biological annotation.

Examine the feature table of the record and verify that the prolactin gene is not annotated there.

Clearly, you could not have found this record using a text search for prolactin. However, in this case, now that there is an assembled zebrafish genome, you could easily have found this BAC clone as a part of the assembly. The “Master” link on the prolactin mRNA “Links” menu leads to the contig that contains this BAC. You could also find the assembly by following the link to Map viewer and adjusting the “Maps & Options” so that the “contig” and “component” maps are displayed. BX511021 appears as one of the components.



Use Splign to perform the same operations.

Splign is the tool currently used at NCBI to make the mRNA to genomic alignments.

Protein and Structures

Example 1: Zebrafish prolactin

Display the Links menu from the zebrafish prolactin mRNA (NM_181437) from the Nucleotides example and follow the link to the protein database.

From the protein record (NP_852102) we can easily find homologs in other species and a structure model for the zebrafish protein. There are generally three options here: running BLAST with the sequence directly, using the pre-computed BLAST results (BLink), or linking to HomoloGene (not available for this record).

Format: [GenPept](#) [FASTA](#) [Graphics](#) [More Formats](#) [Download](#) [Save](#) [Links](#)

NCBI Reference Sequence: NP_852102.2

prolactin [Danio rerio]

[Comment](#) [Features](#) [Sequence](#)

LOCUS NP_852102 210 aa linear VRT 26-JUL-2009

DEFINITION prolactin [Danio rerio].

ACCESSION NP_852102 XP_001331444

VERSION NP_852102.2 GI:127138931

DBSOURCE REFSEQ: accession [NM_181437.3](#)

KEYWORDS .

SOURCE Danio rerio (zebrafish)

ORGANISM [Danio rerio](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Ostariophysi; Cypriniformes; Cyprinidae; Danio.

REFERENCE 1 (residues 1 to 210)

AUTHORS Nguyen,N. and Zhu,Y.

TITLE Prolactin functions as a survival factor during zebrafish embryogenesis

JOURNAL Comp. Biochem. Physiol., Part A Mol. Integr. Physiol. 153 (1), 88-93 (2009)

PUBMED [19032987](#)

REMARK GeneRIF: Prolactin acts as a survival factor during zebrafish embryogenesis.

REFERENCE 2 (residues 1 to 210)

AUTHORS Dutta,S., Dietrich,J.E., Westerfield,M. and Varga,Z.M.

TITLE Notch signaling regulates endocrine cell specification in the zebrafish anterior pituitary

JOURNAL Dev. Biol. 319 (2), 248-257 (2008)

PUBMED [18534570](#)

REFERENCE 3 (residues 1 to 210)

AUTHORS Liu,N.A., Ren,M., Song,J., Rios,Y., Wawrowsky,K., Ben-Shlomo,A., Lin,S. and Melmed,S.

TITLE In vivo time-lapse imaging delineates the zebrafish pituitary proopiomelanocortin lineage boundary regulated by FGF3 signal

JOURNAL Dev. Biol. 319 (2), 192-200 (2008)

PUBMED [18514643](#)

Change Region Show

Customize View

Sequence Analysis Tools

- ▶ BLAST Sequence
- ▶ Conserved Domains

Articles about the protein

- ▶ Prolactin functions as a survival factor during zebrafish embryogenesis [Gen Comp Physiol A Mol Integr Physiol]
- ▶ The effects of the maternal prolactin hormone on zebrafish embryogenesis [Gen Comp Physiol A Mol Integr Physiol]
- ▶ Period2 expression in the developing zebrafish pituitary [Dev Biol]

Identical Proteins for

- ▶ Prolactin [Danio rerio]

RefSeq mRNA

See reference mRNA sequence for gene (NM_181437.3).

More about the prl gene

Also Known As: zgc:110500

BLink

Conserved Domains

Full text in PMC

Gene

Genome Project

Identical Proteins

Nucleotide

PubMed (RefSeq)

PubMed (Weighted)

Related Structure

UniGene

Related Sequences

Domain Relatives

Genome

Map Viewer

PubMed

Taxonomy

Using BLink to find homologs

The BLink output provides direct access to sequence similarity results that are equivalent to BLAST search against the default (nr) protein database.

Click on the BLink link from the zebrafish prolactin protein record (NP_852102).

BLINK precomputed BLAST

Home Taxonomy Report Multiple Alignment **Blast** Help

My NCBI [Sign In] [Register]

Pre-computed BLAST results for: [gi|127138931|ref|NP_852102.2](#) prolactin [Danio rerio]

Matching gis: [62185655](#)

Total (score > 100) : 1643 hits in 1643 proteins in 293 species

Selected: 1643 hits in 1643 proteins in 293 species Filter: Min Score: 100 |


Other views (Reports): [Taxonomy report](#) [Multiple Alignment](#) [Blast](#)

[Reset all filters](#)


[Choose Display Options](#)

0 Archaea 0 Bacteria **1601** Metazoa 0 Fungi 0 Plants 0 Viruses **42** The Others [reset selection](#)

Results: 1 - 100 [Next Page](#) [Last](#)

% hits 

210 aa

blink 

SCORE	ACCESSION	Length	Protein Description
1082	AAH92358	210	Prolactin [Danio rerio]
1077	AAN08916	210	prolactin [Danio rerio]
1008	ABY28339	210	prolactin [Ctenopharyngodon idella]
1008	ABU49656	210	prolactin [Ctenopharyngodon idella]
1004	P29235	210	RecName: Full=Prolactin; Short=PRL; Flags: Precursor
1004	CAA43383	210	unnamed protein product [Hypophthalmichthys nobilis]
997	ABJ90338	210	prolactin [Tinca tinca]
985	CAA37063	210	prolactin [Cyprinus carpio]
985	P09585	210	RecName: Full=Prolactin; Short=PRL; Flags: Precursor
985	CAA31060	210	prolactin (AA -23 to 187) [Cyprinus carpio]
983	AAB47155	210	prolactin; gfPRL [Carassius auratus]
983	AAB47156	210	prolactin; gfPR [Carassius auratus]
983	AAT74865	210	prolactin [Carassius auratus]
983	P87495	210	RecName: Full=Prolactin; Short=PRL; Flags: Precursor
977	CAA43386	207	prolactin [Hypophthalmichthys molitrix]
977	P35395	207	RecName: Full=Prolactin; Short=PRL; Flags: Precursor
955	ACX31825	209	prolactin [Schizothorax prenanti]
816	ABX38813	212	prolactin [Silurus meridionalis]
811	AAK53436	212	prolactin hormone [Heteropneustes fossilis]
792	AAF82287	212	prolactin [Ictalurus punctatus]
792	P51904	212	RecName: Full=Prolactin; Short=PRL; Flags: Precursor

This output shows the non-redundant hits from a protein-protein BLAST search against nr. Notice that there is often more than one protein in the list from the same species. This can occur because of multiple entries with different sequences for the same protein or because the protein belongs to a family of related proteins— in this case the growth hormone family with several members in each organism.

To make it easier to find the one protein for each organism that has the best BLAST score, open the “Choose Display Options” section, select the “best hits” radio button, and click the “BLINK” button.

▼ **Choose Display Options**

filter hits ☒ best hits ☐ all hits ☐ hide identical

Minimum Hit Score **Maximum Hit Score**

New Search By GI **GO** **Items per page**

BLINK

Parameters have been changed. Please, press BLINK button to update the view.

The output now shows one protein from each organism in the list identified by the species name.

Archaea
 Bacteria
 Metazoa
 Fungi
 Plants
 Viruses
 The Others
 [reset selection](#)

Results: 1 - 100 [Next Page](#) [Last](#)

% hits [reset selection](#)

210 aa

blink

SCORE	ACCESSION	N	Tax
1077	AAN08916	11	Danio rerio
1008	ABY28339	7	Ctenopharyngodon idella
1004	P29235	1	Hypophthalmichthys nobilis
997	ABJ90338	2	Tinca tinca
985	CAA37063	10	Cyprinus carpio
983	AAB47155	8	Carassius auratus
977	CAA43386	2	Hypophthalmichthys molitrix
955	ACX31825	4	Schizothorax prenanti
816	ABX38813	1	Silurus meridionalis
811	AAK53436	3	Heteropneustes fossilis
792	AAF82287	6	Ictalurus punctatus
734	P21993	5	Oncorhynchus mykiss
734	P48096	5	Salmo salar
732	CAA45407	7	Oncorhynchus keta
731	AAA51434	2	Coregonus autumnalis
729	Q91364	6	Oncorhynchus tshawytscha
713	Q72ZV3	4	Anguilla japonica
705	BAG72203	2	Plecoglossus altivelis
700	CAA48902	4	Anguilla anguilla
670	AAD15746	4	Paralichthys olivaceus
644	AAO11695	5	Epinephelus coioides
643	BAE45636	3	Thunnus thynnus

◆	—	273	Q28632	2	<i>Oryctolagus cuniculus</i>
◆	—	273	AAQ76548	10	<i>Equus caballus</i>
◆	—	272	CAH05020	2	<i>Nycticebus pygmaeus</i>
◆	—	272	AAX99163	1	<i>Trachypithecus poliocephalus</i>
◆	—	271	P33089	3	<i>Balaenoptera borealis</i>
◆	—	271	ABQ43471	2	<i>Papio anubis</i>
◆	—	270	AAN78320	2	<i>Ailuropoda melanoleuca</i>
◆	—	269	P55151	15	<i>Macaca mulatta</i>
◆	—	269	XP_001171378	48	<i>Pan troglodytes</i>
◆	—	269	AAV17320	7	<i>Nomascus leucogenys</i>
◆	◀	269	XP_545363	4	<i>Canis lupus familiaris</i>
◆	—	268	AAX99162	4	<i>Pithecia pithecia</i>
◆	—	267	NP_000939	47	<i>Homo sapiens</i>
◆	—	267	AAX42634	13	synthetic construct
◆	—	266	P01238	13	<i>Sus scrofa</i>
◆	—	265	CAH05221	4	<i>Callithrix jacchus</i>
◆	—	253	gi 224452	3	<i>Balaenoptera physalus</i>
◆	—	243	ACC59788	11	<i>Capra hircus</i>
◆	—	240	AAI48125	35	<i>Bos taurus</i>
◆	—	240	ABY60851	6	<i>Bubalus bubalis</i>
◆	—	239	Q6UC74	3	<i>Cervus elaphus</i>
◆	—	238	P01240	43	<i>Ovis aries</i>
◆	—	237	CAA24561	54	<i>Rattus norvegicus</i>
◆	—	236	gi 223892	62	<i>Mus musculus</i>
◆	—	232	P37884	3	<i>Mesocricetus auratus</i>
◆	—	232	ACQ73167	1	<i>Callorhinchus milii</i>
◆	—	218	AAD53180	1	<i>Microtus montebelli</i>

The most similar sequences are other fish prolactins, but further down the list are hits to mammalian prolactins including human, mouse and rat. For the mouse human and rat sequences in this output, the protein sequence that is shown is an arbitrarily chosen member of a non-redundant set. The NCBI reference sequence (RefSeq) is usually the most useful record from one of these redundant sets. You can adjust the Display Options so that only the search results against the RefSeq database are shown.

▼ [Choose Display Options](#)

filter hits ☐ best hits ☒ all hits ☐ hide identical ☒ keep only REFSEQ

Minimum Hit Score: 100 Maximum Hit Score:

New Search By GI: GO Items per page: 100

Parameters have been changed. Please, press BLINK button to update the view.

Re-set the radio button to “all hits”, select “REFSEQ” from the “Keep only” pull-down list, and click the “BLINK” button.

0 Archaea 0 Bacteria 168 Metazoa 0 Fungi 0 Plants 0 Viruses 0 The Others [reset selection](#)

Results: 1 - 100 [Next Page](#) [Last](#)

% hits [reset selection](#)

210 aa

blink

SCORE	ACCESSION	Length	Protein Description
734	NP_001118205	210	prolactin precursor [Oncorhynchus mykiss]
734	NP_001117140	210	prolactin precursor [Salmo salar]
562	NP_001072092	213	prolactin [Takifugu rubripes]
306	NP_990797	229	prolactin precursor [Gallus gallus]
290	NP_001036806	229	prolactin precursor [Felis catus]
289	NP_001089341	230	hypothetical protein LOC734391 [Xenopus laevis]
286	NP_001086486	230	prolactin [Xenopus laevis]
286	NP_001093699	230	prolactin [Xenopus (Silurana) tropicalis]
286	NP_001159915	211	prolactin 2 [Xenopus laevis]
277	XP_001513988	227	PREDICTED: similar to preprotein translocase [Ornithorhynchus anatinus]
276	NP_001028166	228	prolactin precursor [Monodelphis domestica]
273	NP_001076144	227	prolactin precursor [Oryctolagus cuniculus]
273	NP_001075365	229	preprolactin precursor [Equus caballus]
269	NP_001040593	227	prolactin precursor [Macaca mulatta]
269	XP_001171378	227	PREDICTED: prolactin isoform 2 [Pan troglodytes]
269	XP_518264	227	PREDICTED: prolactin isoform 3 [Pan troglodytes]
269	XP_545363	259	PREDICTED: similar to Prolactin precursor (PRL) [Canis familiaris]
267	NP_000939	227	prolactin precursor [Homo sapiens]
267	NP_001157030	227	prolactin precursor [Homo sapiens]
263	NP_999091	229	prolactin precursor [Sus scrofa]
241	NP_001156326	226	prolactin 2 [Danio rerio]
240	NP_776378	229	prolactin precursor [Bos taurus]
238	NP_001009306	240	prolactin [Ovis aries]
236	NP_036761	226	prolactin precursor [Rattus norvegicus]

The resulting output provides a list of the best matching RefSeq proteins. The human, mouse and rat prolactin sequences are easily identified. The linked accessions (NP_000939, NP_035294 and NP_036761) will retrieve those records. The linked SCORE will launch BLAST 2 Sequences to compare the zebrafish sequence with the listed one from the other species.

Using Related Structure to find a structure model

The Related Structure shortcut on the Links menu of protein records provides access to BLAST results against the protein sequences from the Structure database and is the fastest way of finding a potential structure for a protein in the database.

Display the Links menu from the zebrafish prolactin protein (NP_852102) and follow the link to Related Structure.

NCBI

Related Structures

Structures related to [gi|127138931|ref|NP_852102.2|]
prolactin [Danio rerio]

2 Low redundancy structures identified

View Low redundancy sequences, sort by BLAST E-value and display as graphic at 20 sequences per page Go

Previous page Jump to page 1 of 1 Next page

Query seq

Protein Families

Superfamilies

Structures

Hormone_1 superfamily

E-value

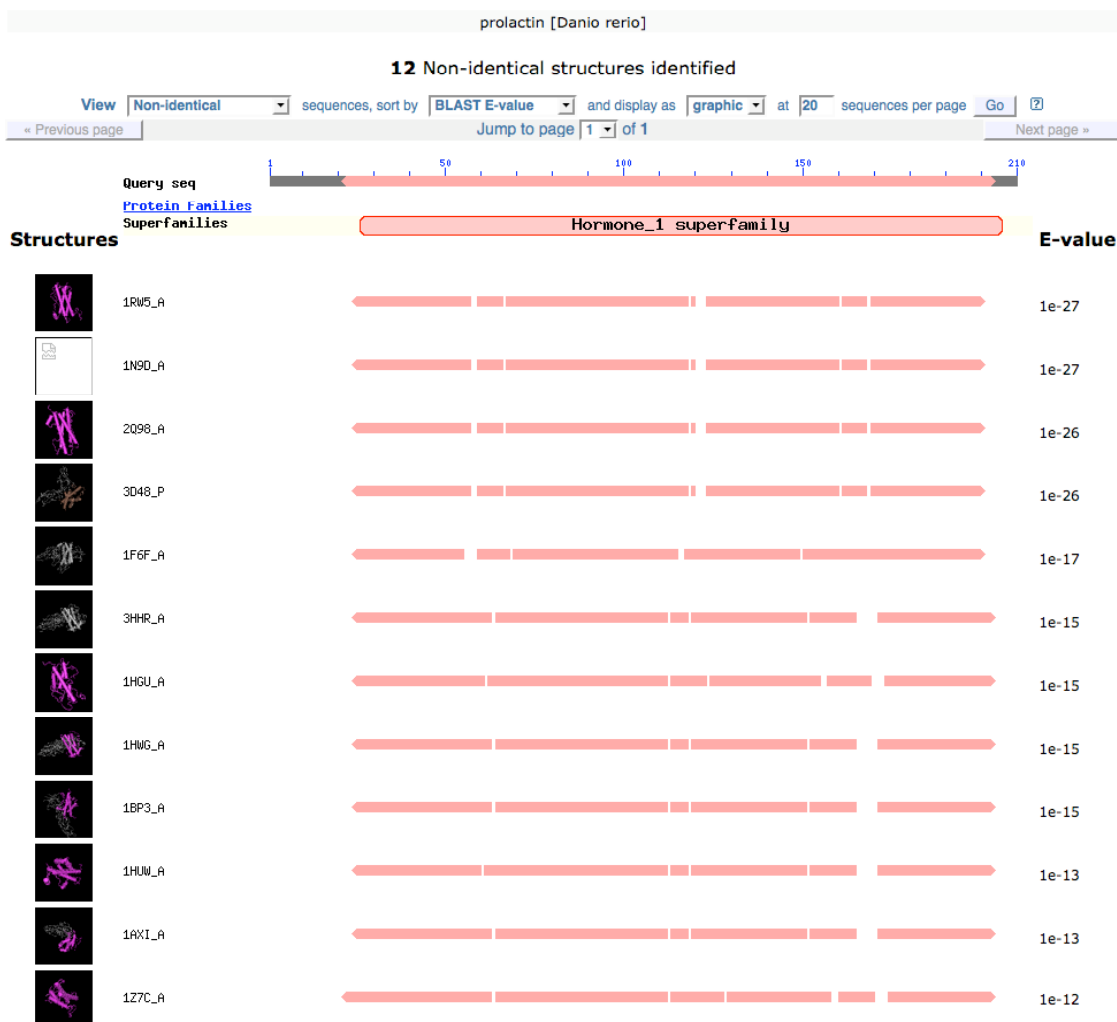
1N90_A 1e-27

2O96_A 1e-26

Previous page Jump to page 1 of 1 Next page

The Related Structure link shows the BLAST alignments of proteins with solved structures.

Increase the number of structure shown by changing the “View” pull-down list from the default “Low redundancy” setting to “non-identical” then click “Go”.



In the expanded list, the first two are human prolactin NMR structures. These are the most similar proteins and would be the best structure models. However these are lower resolution structures than the X-ray crystal structures for the growth hormones that are available. Notice the drop in E-value (significance) from prolactin (1e-27) to the growth hormone entries (1e-15 to 1e-12).

The structure 1BP3 is an X-ray crystal structure of the human growth hormone in a complex with the extra-cellular portion of the prolactin receptor.

Follow the linked identifier [1BP3_A](#) to the structure summary for 1BP3.

The structure contains two chains; the A chain, the growth hormone, and the B chain, the extracellular domains of the prolactin receptor.

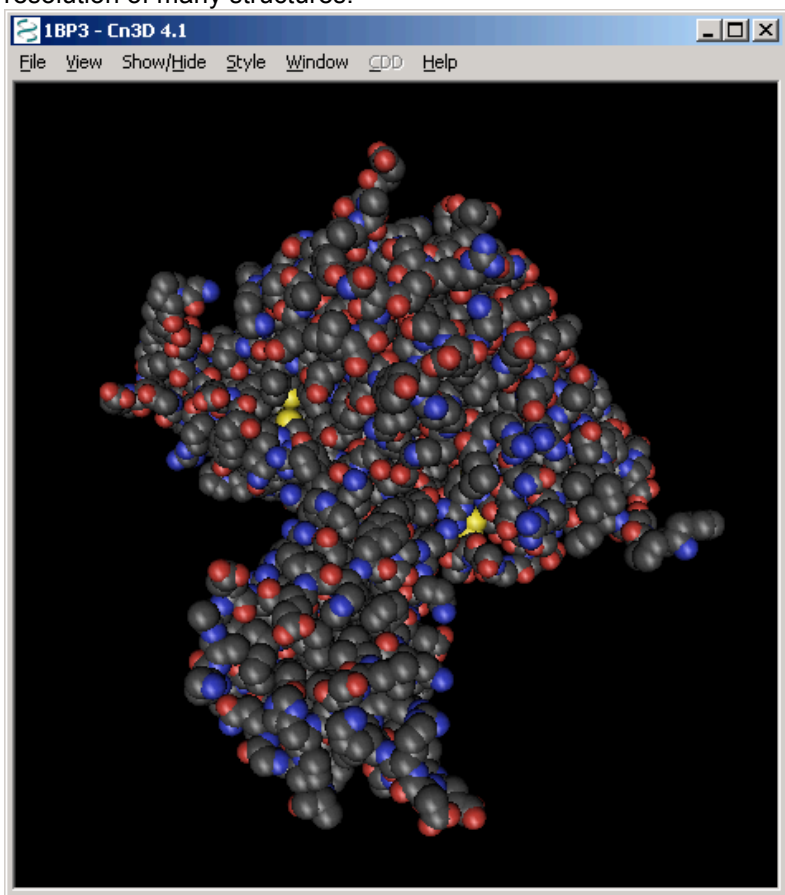
Click the “View 3D structure” button to display the structure in Cn3D.

(Cn3D must be installed on your computer first. If you are attending an NCBI workshop, Cn3D should already have been installed. You can install Cn3D on your own computer by following the instructions linked to "Download Cn3D".)

The structure is displayed showing only the alpha carbon backbone. It is colored by secondary structure, and the secondary structure regions are indicated by special objects. The alpha helices are indicated with green cylindrical arrows pointing in the C-terminal direction, the beta strands are indicated by flat tan arrows. You change the rendering of the structure through the Style menu of the viewer.

Use the Style menu and the Rendering Shortcuts to change to Space Fill. Then use the Style Coloring Shortcuts to color by Element.

This now more closely resembles a molecular model of the entire complex including the amino acid side chains. Notice that all of the elements are represented except for hydrogen (carbon = black, oxygen = red, nitrogen = blue, sulfur = yellow). Hydrogen atoms are below the limit of resolution of many structures.



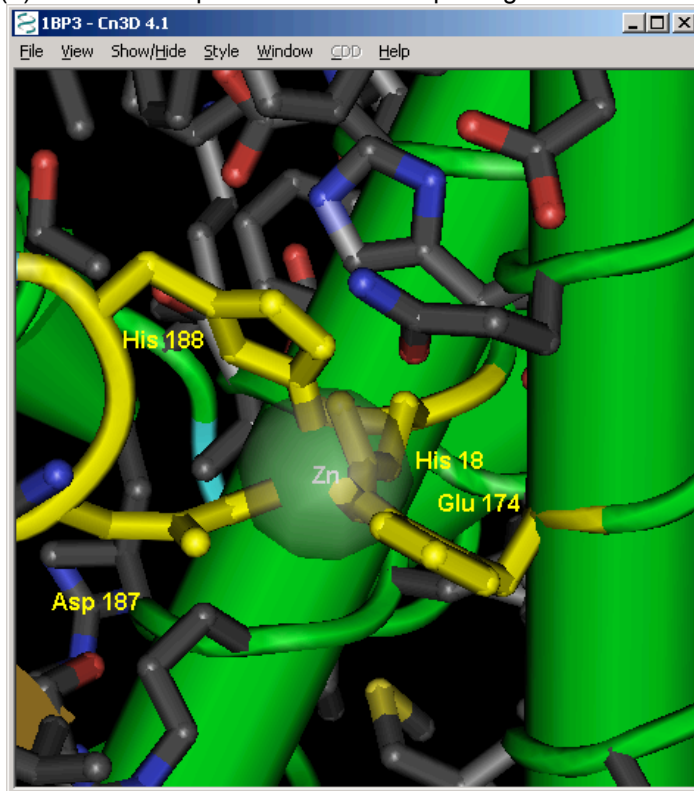
Restore the earlier rendering and coloring by setting Style:Rendering Shortcuts:Worms and Style:Coloring Shortcuts:Secondary Structure.

Hold the mouse button down and rotate the structure so that the zinc ion is visible. Use the View menu and Zoom In to get a closer view of the region surrounding the zinc ion.

If the zinc ion gets off-center you can hold the shift key down and drag the structure with mouse while holding button down. You can turn on the side chains to see which ones are coordinating the zinc ion.

- **Use the Style:Rendering Shortcuts:Toggle Sidechains to turn on the amino acid side chains.**
- **Use Style: Edit Global Style to display the Global Style menu and change the rendering of the Protein side chains to Tubes to make them easier to see.**
- **You can highlight the four amino acids that are making contact with the zinc by double clicking on the residues in the structure viewer.**
- **Notice that the residue highlighted in the structure are also highlighted in the Sequence/Alignment Viewer window**

There is a histidine (h) and a glutamate (e) from the hormone and an aspartate (d) and a histidine (h) from the receptor involved in complexing the zinc ion.



Example 2: Human MutL Homolog 1

MLH1 is the product of a well-known human disease gene that is mutated in some heritable cancer syndromes.

Use the global query to retrieve the Swiss-Prot record for human DNA mismatch repair protein MLH1. To save time, you can retrieve it directly using the accession, P40692.

Of course, you could perform a global text search for mlh1, retrieve the protein results, then use the "Limits" tab as you did in Example 1 above to obtain precise results.

P40692 is a record imported from the Swiss-Prot database. Swiss-Prot is a smaller database of highly informative protein records. Many of them are equivalent to review articles on a particular protein. The present record has a large amount of information on the biology of MLH1 including a large list of polymorphisms.

Examine the FEATURES table of the record and locate several of the polymorphisms in the first 50 residues of the protein.

FEATURES	Location/Qualifiers
source	1..756 /organism="Homo sapiens" /db_xref="taxon: 9606 "
gene	1..756 /gene="MLH1" /note="synonym: COCA2"
Protein	1..756 /gene="MLH1" /product="DNA mismatch repair protein Mlh1"
Region	1..756 /gene="MLH1" /region_name="Mature chain" /experiment="experimental evidence, no additional details recorded" /note="DNA mismatch repair protein Mlh1." /FTId=PRO_0000178000."
Region	8..>575 /gene="MLH1" /region_name="MutL" /note="DNA mismatch repair enzyme (predicted ATPase) [DNA replication, recombination, and repair]; COG0323" /db_xref="CDD: 30671 "
Region	18 /gene="MLH1" /region_name="Variant" /experiment="experimental evidence, no additional details recorded" /note="R -> C (in HNPCC2). /FTId=VAR_022663."
Region	28 /gene="MLH1" /region_name="Variant"

```

/experiment="experimental evidence, no
additional details recorded"
/note="P -> L (in HNPCC2). /FTId=VAR_004433."
31..122
/region_name="HATPase_c"
/note="Histidine kinase-like ATPases; This family
includes several ATP-binding proteins for
example: histidine kinase, DNA gyrase B,
topoisomerases, heat shock protein
HSP90, phytochrome-like ATPases and DNA mismatch
repair proteins; cd00075"
/db_xref="CDD:28956"
32
/region_name="Variant"
/experiment="experimental evidence, no
additional details recorded"
/note="I -> V (in
dbSNP:rs2020872). /FTId=VAR_014876."
order(34,38,41,61,63,65,67..68,101..104,115,
117,122)
/site_type="other"
/note="ATP binding site"
/db_xref="CDD:28956"
35
/region_name="Variant"
/experiment="experimental evidence, no
additional details recorded"
/note="M -> R (in HNPCC2). /FTId=VAR_004434."
37
/region_name="Variant"
/experiment="experimental evidence, no
additional details recorded"
/note="E -> ELNH (in endometrial cancer; somatic).
/FTId=VAR_004435."
38
/site_type="other"
/note="Mg2+ binding site"
/db_xref="CDD:28956"
44
/region_name="Variant"
/experiment="experimental evidence, no
additional details
recorded"
/note="S -> F (in HNPCC2; the equivalent

```

```
substitution in yeast causes loss of function in  
a mismatch repair assay).  
/FTId=VAR_004436."
```

Several of these polymorphisms are annotated with the name of a disease or syndrome, for example, hereditary non-polyposis colorectal cancer type 2 (HNPCC2). There is also a polymorphism at position 32 that is cross-referenced to NCBI's dbSNP. In the following sections, you will use some of the pre-computed Entrez relationships to map these polymorphisms onto a 3D structure.

Links: Related Sequences

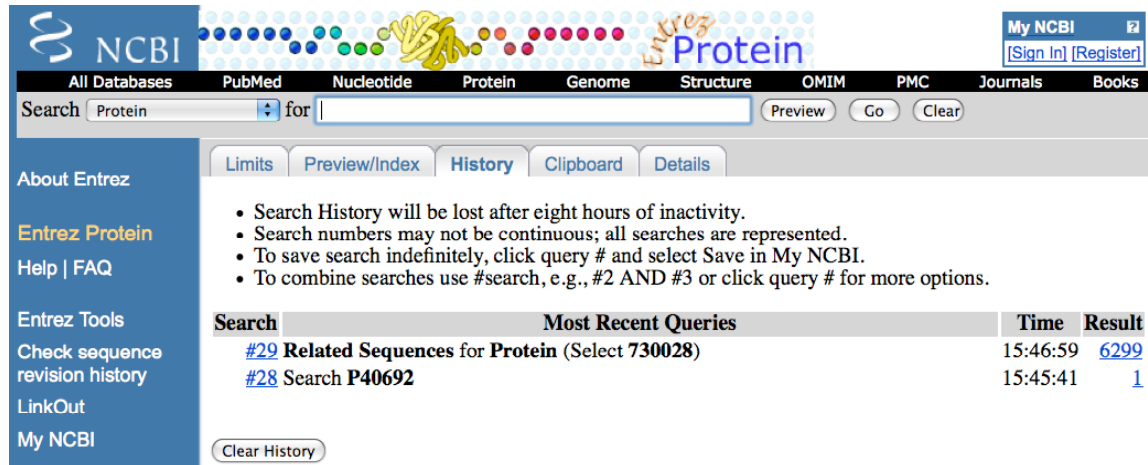
Use the "Links" menu as in Example 1 to display "Related Sequences".

The resulting display is a list of similar sequences arranged in descending order by BLAST score as with the nucleotide "Related Sequences". Unlike the nucleotide "Related Sequences", the protein similarities typically do find sequences from other species. How exactly the protein sequences in this list are related to the sequence in P40692 is not easily seen. Some of these proteins are identical to P40692; some are very similar over the entire length, some share only a domain in common. All that the list tells you is that the sequences are significantly related. Although it isn't obvious, the first several proteins are, in fact, identical sequences. That set includes corresponding records representing this human protein from at least four different sources; Swiss-Prot, PRF, RefSeq and more than one translation of a GenBank/EMBL/DDBJ sequence. The inclusion of records from outside protein databases plus our own RefSeq database results in a high degree of redundancy at the sequence level in the protein data. The records themselves are not redundant, however, since the annotation on the records is different. When creating a BLAST database and for BLink, identical sequences are represented as a single sequence. The non-redundant database is about 50% smaller than the entire Entrez protein database.

Change the "Display" drop-down menu to show 500 records. Scroll through the list to see records from other species.

There are proteins in the list from a wide range of taxa: bacteria, green plants, protozoa, multicellular animals. Although the distance of a particular protein from the top of the list appears to approximate the evolutionary distance from human, keep in mind that some proteins in the list are fragments and may have low scores simply because they are short. You modify the search to find all of the proteins from a particular taxon through the "History" tab.

Click on the "History" tab.



NCBI Entrez Protein

My NCBI [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search Protein for [] Preview Go Clear

About Entrez

Entrez Protein

Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

My NCBI

Limits Preview/Index History Clipboard Details

- Search History will be lost after eight hours of inactivity.
- Search numbers may not be continuous; all searches are represented.
- To save search indefinitely, click query # and select Save in My NCBI.
- To combine searches use #search, e.g., #2 AND #3 or click query # for more options.

Search	Most Recent Queries	Time	Result
#29	Related Sequences for Protein (Select 730028)	15:46:59	6299
#28	Search P40692	15:45:41	1

Clear History

This is the protein search history that is maintained on our Web server. You can combine the entries in your history with other searches. For example, you can combine the entry for related proteins, called Protein Neighbors, with an organism query.

Type the number of the entry in your history for the Protein Neighbors in the search box followed by an organism search for mouse. For example

#29 AND mouse[Organism]

You will need to turn Limits off if you used them previously.

Then run the search.

There are several proteins from mouse in the related sequences that are now displayed. Since the related sequences search is combined with another Entrez search, the sorting order is lost. The mouse proteins are listed in arbitrary order, not by their BLAST score with the human MLH1. The BLink option that used in Example 1 and below makes it much easier to find homologs in other species. It also allows you to see alignments themselves.

Links: Finding a related structure

Previously you saw that there are a number of sources that contribute to the protein database. One source is the Protein Databank (PDB). PDB is a database of 3D biomolecular structures. NCBI imports these structures and makes them available in the Entrez system as the Structure database. In addition, protein sequences are extracted from the structures and entries are created in the protein database. This makes it easy to find a structure for a particular protein or a homolog if one exists. Several related proteins in the MLH1 example are PDB entries and have links to the structure database.

Use the browser “Back” button or the “History” tab to return to the list of related sequences to P40692. Use the “Display” drop down menu to select “Structure Links.” The page will automatically refresh.

NCBI Entrez Protein

Search Protein for Go Clear

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort By Send to

All: 62 Items

OMIM Links
BioAssay Links
BioAssay by target
BioAssay by target, identical sequence
PubChem Compound Links
PubChem Substance Links
Peptidome Links
PMC Links
PopSet Links
Protein (UniProtKB)
Protein (RefSeq)
Protein Cluster Links
PubMed Links
PubMed (RefSeq) Links
PubMed (Weighted) Links
SNP Links
Gene Genotype Links
Structure Links
Taxonomy Links
UniGene Links

Related Structures: 3447

Page 1 of 315 Next

protein Mlh1; AltName: Full=MutL protein

lypD type 2 (E. coli) [Homo sapiens]

756 aa protein
AAE56022.1 GI:14107168

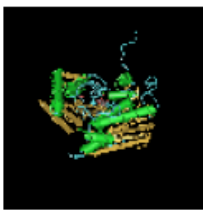
Top Organisms [Tree]
marine metagenome (955)
Homo sapiens (142)
Escherichia coli (131)
Staphylococcus aureus (106)
Streptococcus pneumoniae (99)
All other taxa (4921)
More...

Recent activity

The new set of results that is displayed contains structure records. Notice that the graphic at the top of the page has changed, and you are now in the Entrez structure database. As with the previous example, the sorting order is lost. Several of these are structures of bacterial DNA mismatch repair proteins.

8: 1B63

[Related Structures](#), [Literature](#), [Domains](#), [Chemicals](#)



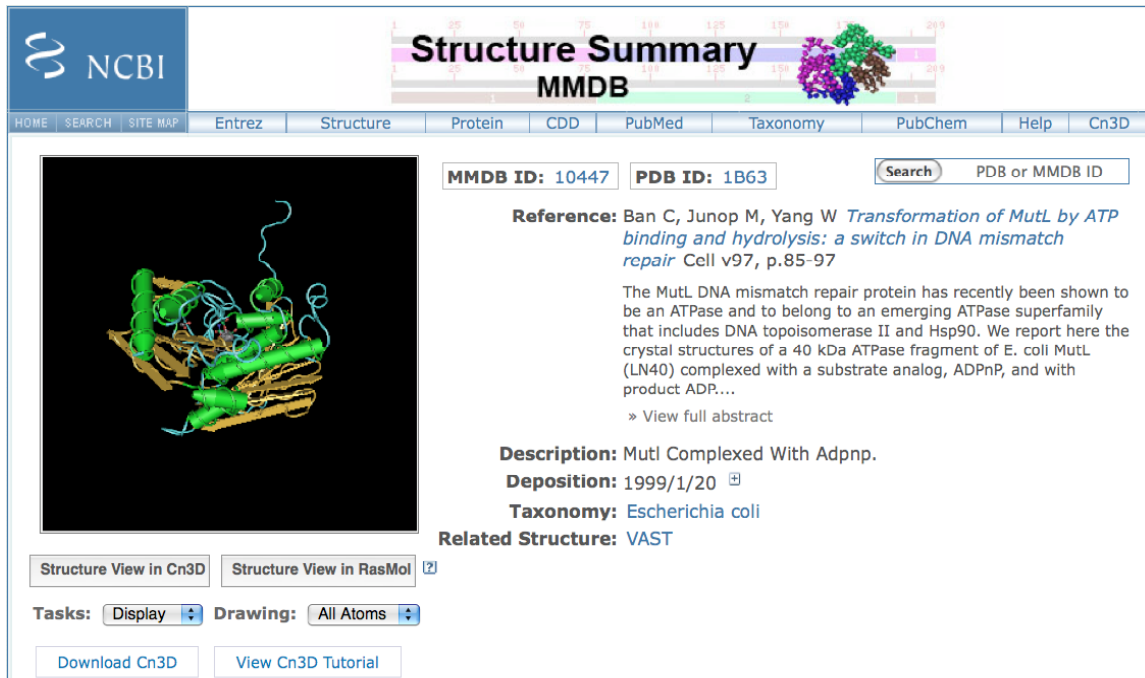
MutL Complexed With Adpnp [Dna Mismatch Repair]

Taxonomy: [Escherichia coli](#)

Proteins: 1; Chemicals: 3

modified: 2007/10/06; MMDB ID: 10447

Retrieve the structure summary for 1B63 by clicking on the linked structure thumbnail image.



NCBI

Structure Summary MMDb

HOME | SEARCH | SITE MAP | Entrez | Structure | Protein | CDD | PubMed | Taxonomy | PubChem | Help | Cn3D

MMDb ID: 10447 PDB ID: 1B63 Search PDB or MMDb ID

Reference: Ban C, Junop M, Yang W *Transformation of MutL by ATP binding and hydrolysis: a switch in DNA mismatch repair* Cell v97, p.85-97

The MutL DNA mismatch repair protein has recently been shown to be an ATPase and to belong to an emerging ATPase superfamily that includes DNA topoisomerase II and Hsp90. We report here the crystal structures of a 40 kDa ATPase fragment of *E. coli* MutL (LN40) complexed with a substrate analog, ADPnp, and with product ADP....

» View full abstract

Description: Mutl Complexed With Adpnp.
Deposition: 1999/1/20
Taxonomy: *Escherichia coli*
Related Structure: VAST

Structure View in Cn3D Structure View in RasMol

Tasks: Display Drawing: All Atoms

Download Cn3D View Cn3D Tutorial

The structure summary page shows a graphic representing the biomolecular chains in the record with the 3D domains and conserved domains mapped onto the chain.

Display the structure by clicking the image of the structure in the summary.

In order to display the structure, you will need to have the NCBI structure viewer, Cn3D installed. If the viewer is not already installed, follow the hyperlink labeled “Get Cn3D” and follow the instructions to install Cn3D.

The 1B63 record is the X-ray crystal structure of the N-terminal portion of the MutL DNA mismatch repair protein from *E. coli*. The default display in Cn3D shows the alpha carbon backbone of the protein colored by the type of secondary structure; alpha helices are green, beta strands are tan, and random coil is blue. There are also 3D objects representing the helices and strands. You can rotate the structure by dragging it with the mouse pointer while holding down the left mouse button. Holding the Shift key down will allow you to move the entire structure by dragging it with the mouse pointer.

You can modify the way the structure is rendered through the “Style” menu of the viewer.

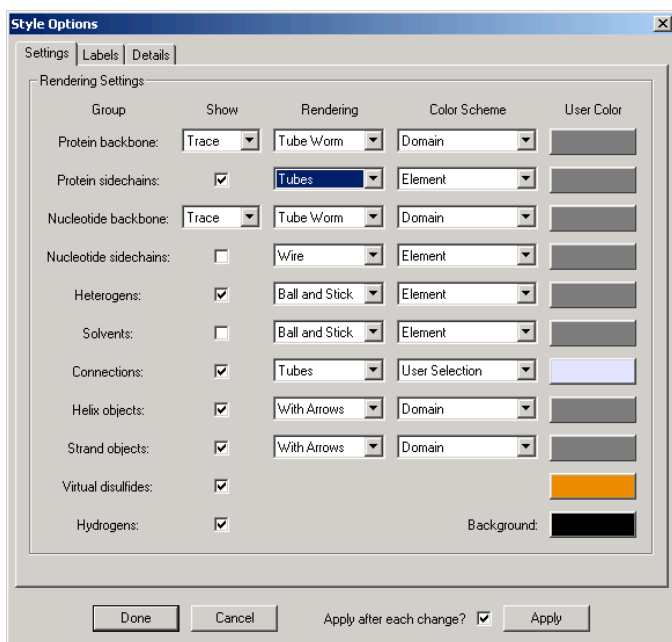
Use the “Coloring shortcuts” on the “Style” menu to color by Domain.

The color scheme matches that on the structure summary Web page. The purple domain corresponds to the 3D domain also identified as a histidine kinase-like ATPase domain. This domain contains many of the protein polymorphisms associated with disease.

Use the “View” menu to zoom in to the ATPase domain.

An ATP analog is co-crystallized in this domain. Oxygen atoms on the three phosphates of the ATP analog make close contact with a magnesium ion. An amino acid side chain completes the coordination sphere of this metal ion. You can turn on the protein side chains to identify this residue.

Now, use the “Style” menu on the viewer to open the “Global” style dialog box.



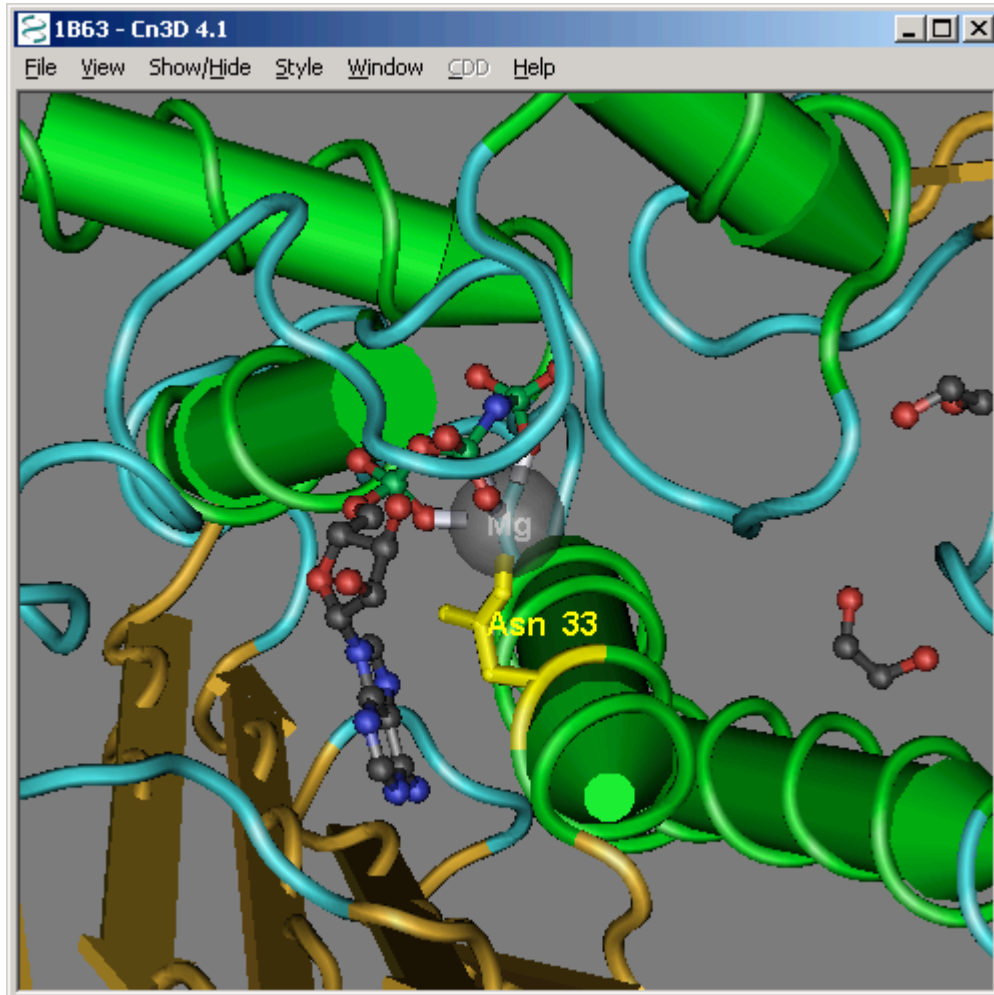
Turn on the protein side chains by checking the box on that line.

You can alter the way the side chains are rendered using the corresponding drop down menu.

Press the “Done” button to close the “Style Options” dialog box.

Zoom in to the region of the protein near the magnesium ion and find the side chain that makes a contact with the metal ion. Double click on this residue to highlight it.

You can identify the residue and its position by looking at the residue now highlighted in yellow in the sequence viewer.



BLink: non-redundant protein neighbors

Use the global query on the NCBI homepage or the Protein database to retrieve P40692 again and follow the BLink link in from the Links menu.

BLink provides a way of viewing related sequences that is more like a standard protein BLAST output. The source database for BLink is essentially the BLAST non-redundant protein database. This is the Entrez set with the biologically uninteresting patent sequences removed. At the top of the page, is a list of the gi number of sequences in the protein database that are identical to P40692. The graphic alignment shows the regions of the proteins that align to the query. The hyperlinked BLAST score shows the detailed alignment between the two proteins.

Follow the instructions for Example 1 to generate the “Best Hits” display.

In many cases there may be more than one protein in the display from the same species. Sometimes this is because of the presence of paralogous proteins or because there may be

differences in the sequences from different sources for the same protein. You can easily identify the best protein match from the tomato, *Solanum lycopersicum*.

Click on the BLAST score on the line containing the best tomato (*Solanum lycopersicum*) protein.

The new window shows the BLAST 2 Sequences alignment between the human MLH1 and the best match in tomato. This is a highly significant local alignment that extends nearly the entire length of both proteins.

As in Example 1, you can use the Display options “Keep only” drop down menu to limit to various subsets of the protein data, for example PDB to find structures as we did with the Entrez related proteins. Another way of finding structures for related proteins is through the “Related structure” shortcut on the protein links menu.

Using Related Structures to find a structural model

The “Related structure” shortcut on the protein links menu provides a simple way to find related structures.

Retrieve the protein record P40692 and follow the link to Related Structure as in Example 1.

Format: [GenPept](#) [FASTA](#) [Graphics](#) [More Formats](#) [Download](#) [Save](#) [Links](#)

★ Try the [Graphics report](#) for a more informative view of the biological features.

Swiss-Prot: P40692.1

RecName: Full=DNA mismatch repair protein Mlh1; AltName: Full=MutL protein homolog 1

[Comment](#) [Features](#) [Sequence](#)

LOCUS P40692 756 aa linear PRI 13-OCT-2009
 DEFINITION RecName: Full=DNA mismatch repair protein Mlh1; AltName: Full=MutL protein homolog 1.
 ACCESSION P40692
 VERSION P40692.1 GI:730028
 DBSOURCE UniProtKB: locus MLH1_HUMAN, accession [P40692](#);
 class: standard.
 created: Feb 1, 1995.
 sequence updated: Feb 1, 1995.
 annotation updated: Oct 13, 2009.
 xrefs: [U07343.1](#), [AAC50285.1](#), [U07418.1](#), [AAA17374.1](#), [U40978.1](#), [AAA82079.1](#), [U40960.1](#), [U40961.1](#), [U40962.1](#), [U40963.1](#), [U40964.1](#), [U40965.1](#), [U40966.1](#), [U40967.1](#), [U40968.1](#), [U40969.1](#), [U40970.1](#), [U40971.1](#), [U40972.1](#), [U40973.1](#), [U40974.1](#), [U40975.1](#), [U40976.1](#), [U40977.1](#), [U17857.1](#), [AAA85687.1](#), [U17839.1](#), [U17840.1](#), [U17841.1](#), [U17842.1](#), [U17843.1](#), [U17844.1](#), [U17845.1](#), [U17846.1](#), [U17847.1](#), [U17848.1](#), [U17849.1](#), [U17851.1](#), [U17852.1](#), [U17853.1](#), [U17854.1](#), [U17855.1](#), [U17856.1](#), [AY217549.1](#), [AAO22994.1](#), [BC006850.1](#), [AAH06850.1](#), [S43085](#), [NP_000240.1](#)
 xrefs (non-sequence databases): IPI:IP100029754, UniGene:[Hs.195364](#), HSSP:[P23367](#), DIP:DIP:27601N, IntAct:P40692, STRING:P40692, PhosphoSite:P40692, PRIDE:P40692, Ensembl:ENST00000231790, Ensembl:ENST00000383761, Ensembl:ENST00000396438, Ensembl:ENST00000396447, Ensembl:ENST00000413212,

Change Region Shown
Customize View

Sequence Analysis Tools
 ▶ BLAST Sequence
 ▶ Conserved Domains

Articles about the MLH1
 ▶ Cancer risk in a cohort of a single mismatch repair protein (RefSeq)
 ▶ A MLH1 polymorphism and cancer risk (Cancer Gene)
 ▶ MLH1 protects from recombination by the histone deacetylase

Identical Proteins for P40692
 ▶ Sequence 20 from patent
 ▶ unnamed protein product
 ▶ mutL homolog 1, colon

Links
 BLink
 Conserved Domains
 BioSystems
 Full text in PMC
 Gene
 Gene Genotype
 GeneView in dbSNP
 Identical Proteins
 Protein (RefSeq)
 PubMed (Weighted)
 Related Structure
 Related Sequences
 Domain Relatives
 OMIM
 PubMed
 Taxonomy
 LinkOut

The related structures display provides an output similar to BLink, providing a BLAST output sorted by similarity with access to the alignments. The thumbnail image of structures (1B63 A) on the left hand side link directly to the structure summary. The pink alignment graphic links to a page with the sequence alignment that allows loading the alignment display in Cn3D.

Click on the “View structure and alignment in Cn3D” button.

The resulting Cn3D display now shows the 1B63 structure colored by sequence conservation from the alignment of the human MLH1 (bottom sequence) and the N-terminal sequence region of MutL (top sequence). You can use the sequence alignment to map the human residues onto the *E. coli* protein structure. In other words, the human protein is assumed to fold up into a very similar structure; the sequence alignment is used as a proxy for the structural alignment. This is reasonable as long as the proteins are similar at the sequence level. You can confirm the validity of this to some extent by verifying that structurally and functionally significant residues in the structure line up with corresponding residues in the aligned protein sequences.

Manipulate the structure in the viewer and use the view menu on the viewer to zoom in to the ATP binding site residue; the asparagine (n) at position 33 of the structure. Verify that this residue is aligned with an asparagine in the human sequence.

You can now look at some of the polymorphisms reported in the FEATURES table of P40692 in the context of the structure of the protein. Notice that the isoleucine to valine change at position 32 of the human protein, which is not reported as associated with human disease, occurs on the side of the helix containing the ATP binding site residue that is away from ATP. In fact, the residue in that position in the *E. coli* protein is a valine. A disease causing polymorphism in the human protein replaces the proline at position 28 of the human protein with a leucine. The proline in this position, which is conserved in *E. coli*, may be important in constraining the turn at the end of the helix.

NCBI Exercises Set 2

NCBI Genomic Resources

Albumins constitute a small family of genes in mammals. The human, mouse and rat genomes, and probably all mammals contain at least four members: albumin, alpha-fetoprotein, afamin (alpha albumin) and the vitamin D binding protein. We will look at various aspects of this gene family in the NCBI genome resources.

UniGene and Gene

UniGene is the best NCBI resource to identify the gene (or suspected gene) that corresponds to a particular database sequence. This is especially true for Expressed Sequence Tags (ESTs) where there may be no annotations on the sequence, but may also be important for other sequences where the annotation may be incomplete or obsolete. Database identifiers for UniGene searches may come from BLAST output or from microarray (hybridization) data. For example, an mRNA that hybridized to the EST sequence with accession number BG618460 was highly expressed in a human liver tumor sample. We can identify this gene using UniGene.

Retrieve BG618460 from the EST database. You can use the search box on the NCBI homepage and retrieve the link to EST on the global query page.

Is there any information on the record indicating what gene this is? The database ads in the right-hand discovery column now immediately reveal that this is an EST from the albumin gene (ALB).

GenBank: BG618460.1

**602645646F1 NIH_MGC_76 Homo sapiens cDNA clone
IMAGE:4767490 5-, mRNA sequence**

IDENTIFIERS

dbEST Id: 8340792
EST name: 602645646F1
GenBank Acc: BG618460
GenBank gi: 13669831

CLONE INFO

Clone Id: IMAGE:4767490 (5')
Plate: LLCM1629 Row: k Column: 11
DNA type: cDNA

PRIMERS

PolyA Tail: Unknown

SEQUENCE

```
GAGCTTTTCTCTCTGTCAACCCACACGCTTTGGCACAATGAAGTGGGTAACCTTTAT
TTCCCTTTCTTTCTCTTTAGCTCGGCTTATCCAGGGGTGTGTTTCGTCGAGATGCACA
CAAGAGTGAAGTTGCTCATCGGTTTAAAGATTGGGAGAAGAAAATTTCAAAGCCTTGGT
GTTGATTGCCCTTGTCTCAGTATCTTCAGCAGTGTCCATTTGAAGATCATGAAAAATTAGT
GAATGAAGTAACTGAATTTGCAAAAACATGTGTTGCTGATGAGTCAGTGAATAATTGTGA
CAATCACTTCATACCTTTTGGAGACAAATATGCACAGTTGCAACTCTCGTGAAAC
CTATGCTGAAACTGGCTGACTGCTGTGCAAAAACAGACCTGAGAGAAATGAATGCTTCT
TGCAACACAAAGATGACANACCACAACTCCCGGATTGGTGAGACCAGAGGTTGATGTGA
TGTCAGCTGCTTTTCATGACAATGAAGAGACATTTTGAACCAATACTTATCTGAAACT
TGCCAGAGACATCCTTACTTTTATGCCCCGGAACCTCTTTCTTTGCTAAAAGGTATAA
AGCTGCTTTTACAGAATGTTGCCAAGCTGCTGATAAAGCCTGCTGCCTGTTGCCAAGCT
CGATGAACCTTCGGGATGACGGAAGGCTTCGCTGTCACACAGAGACTCAGTGTGCCAGTTC
TCCACAACTTTGGAGACAGGAGCTTTTCAAAGCCATGGGCACTAGTTTCGCTGAGCAAG
AAATTCCTCCAAAGGTTGAGTTTGGAGACAGTTCCTCAAGTTAGTGAACGAATCTCAGAGTTC
CACCGGAATGCTTGCCTGGAGATCTGGTCAATGTCTGTGACACGGGGGCTTCAG
AGATTC
```

Quality: High quality sequence stops at base: 719

Entry Created: Apr 17 2001
Last Updated: Apr 18 2001

COMMENTS

Tissue Procurement: CLONETECH Laboratories, Inc.
cDNA Library Preparation: CLONETECH Laboratories, Inc.
cDNA Library Arrayed by: The I.M.A.G.E. Consortium (LLNL)
DNA Sequencing by: Incyte Genomics, Inc.
Clone distribution: MGC clone distribution information can
be found through the I.M.A.G.E. Consortium/LLNL at:
<http://image.llnl.gov>

Sequence Analysis Tools

- ▶ BLAST Sequence
- ▶ Pick Primers

RefSeq mRNA

See the ALB reference mRNA sequence (NM_000477.5) for this EST.

ESTs for the ALB gene

This EST is one of 16904 sequences matched to ALB: Albumin.

More about the ALB gene

Albumin is a soluble, monomeric protein which comprises about one-half of the blood serum protein. Albumin functions primarily as a carrier ...

Also Known As: DKFZp779N1935, PRO0883,...

Homologs of the ALB gene

The ALB gene is conserved in chimpanzee, dog, cow, mouse, rat, and chicken.

Recent activity

All links from this record

- ▶ Taxonomy
- ▶ Map Viewer
- ▶ Traces
- ▶ UniGene

Follow the ad for the Gene database ("More about the ALB gene") to identify the function of this gene and its products.

Go back to and follow ad for UniGene ("ESTs for the ALB gene").

Look at the ESTs in this cluster. How many are there? A pair of ESTs (a 5' and 3' read) that come from the same clone ID are T58928 and T58869. You'll need to display all ESTs and scroll down to see these. Also, identify the RefSeq mRNA in the cluster. You should be able to recognize the RefSeq by the characteristic accession.

Link to the BLAST homepage and use BLAST 2 Sequences to align the 5' and 3' reads to the RefSeq mRNA.

Notice the mismatches that are most likely due to sequencing errors in the ESTs. Expression information is implied by the sources of the cDNA libraries in a particular cluster.

Follow the "Expression profile" mapping link to see a "virtual Northern" display of the counts from this cluster in UniGene libraries.

What library pool shows the highest relative expression of this gene?

Map Viewer

From the “Gene” page for human *Alb*, use the “Links” menu to display this gene in the Map Viewer.

What chromosomal region is this? What maps are displayed? You can click on the map name at the top to learn more about the information displayed for each map. The UniGene map shows the density of EST hits on the genome. Generally the peaks in this histogram highlight the exons of expressed genes. Notice that there are some hits that don't correspond to the exons shown in the gene model on the Genes map. What could these represent?

You may want to use the “Maps and Options” dialog box to remove all except the “Gene” map from the display for easier viewing. The “Maps and Options” can be accessed through the button on the upper right of the maps. Click “OK” once you adjust the maps.

Organism: **Homo sapiens**

[Help](#)

Chromosome: Region Shown:

Available Maps:

Org:

Assembly:

---Sequence Maps---

- ☐ Ab initio
- ☐ Assembly
- ☐ Celera Genes
- ☐ Celera Transcripts
- ☐ Clone
- ☐ Component
- ☐ Contig
- ☐ CpG Island

Maps Displayed (left to right):

☐ Ab initio
☐ ugHs
☐ Ensembl Genes
☐ RefSeq Transcripts
☒ Gene

([R] before map means 'ruler set')

More Options:

☐ Show Connections ☒ Verbose Mode

Compress Map: Auto Compress if > px

Page Length:

Thumbnail View: ☒ default (ideogram) ☐ master

Use the zoom graphic on the left hand side of the map viewer to zoom out and display two other members of the albumin gene family, *AFP* and *AFM*.

Are these in the same orientation?

[Homo sapiens \(human\)](#) [Build 37.1 \(Current\)](#)

[BLAST The Human Genome](#)

Chromosome: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [11](#) [12](#) [13](#) [14](#) [15](#) [16](#) [17](#) [18](#) [19](#) [20](#) [21](#) [22](#) [X](#) [Y](#) [MT](#)

Query: 213[[gene_id](#)] [\[clear\]](#)

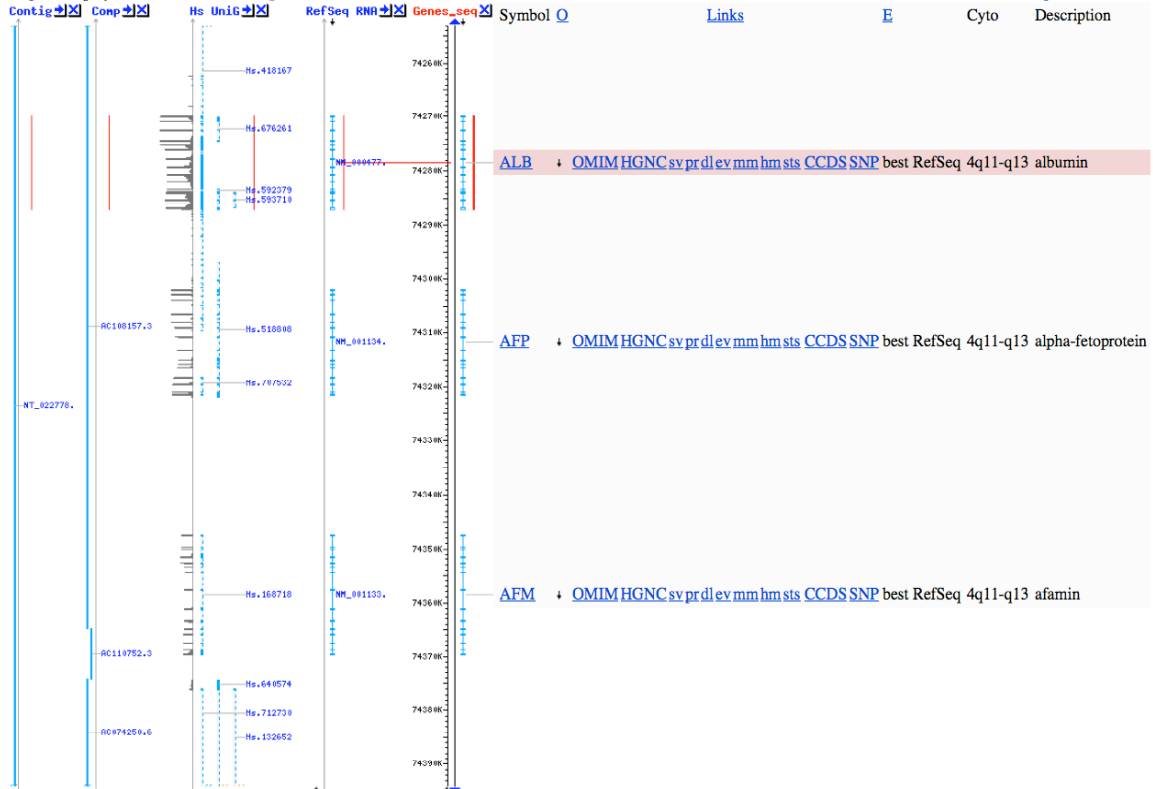
Master Map: [Genes On Sequence](#)

[Summary of Maps](#)

[Maps & Options](#)

Region Displayed: 74,253K-74,394K bp

[Download/View Sequence/Evidence](#)



The fourth member of this small family, the vitamin D binding protein, also called group-specific component (GC), is somewhat removed from these on chromosome 4.

Display the entire region between GC and AFM by typing these symbols in the "Region Shown" boxes on the left-hand-side and pressing the "Go" button.

Use the "Maps and Options" link to add the mouse and rat gene maps to the display.

Organism: Homo sapiens[Help](#)**Chromosome:** **Region Shown:** **Available Maps:**Org:

Conti chimp
CpG I human
Enser mouse
Enser **rat**
GenBank DNA
Gene
Phenotype
RefSeq Transcripts
Repeats

ADD>>

<<REMOVE

Maps Displayed (left to right):

[R][mouse][human:4] Gene
[R][rat][human:4] Gene
[R][human][4] Gene

Move UP

Move DOWN

Make Master/Move to Bottom

Toggle Ruler

([R] before map means 'ruler set')

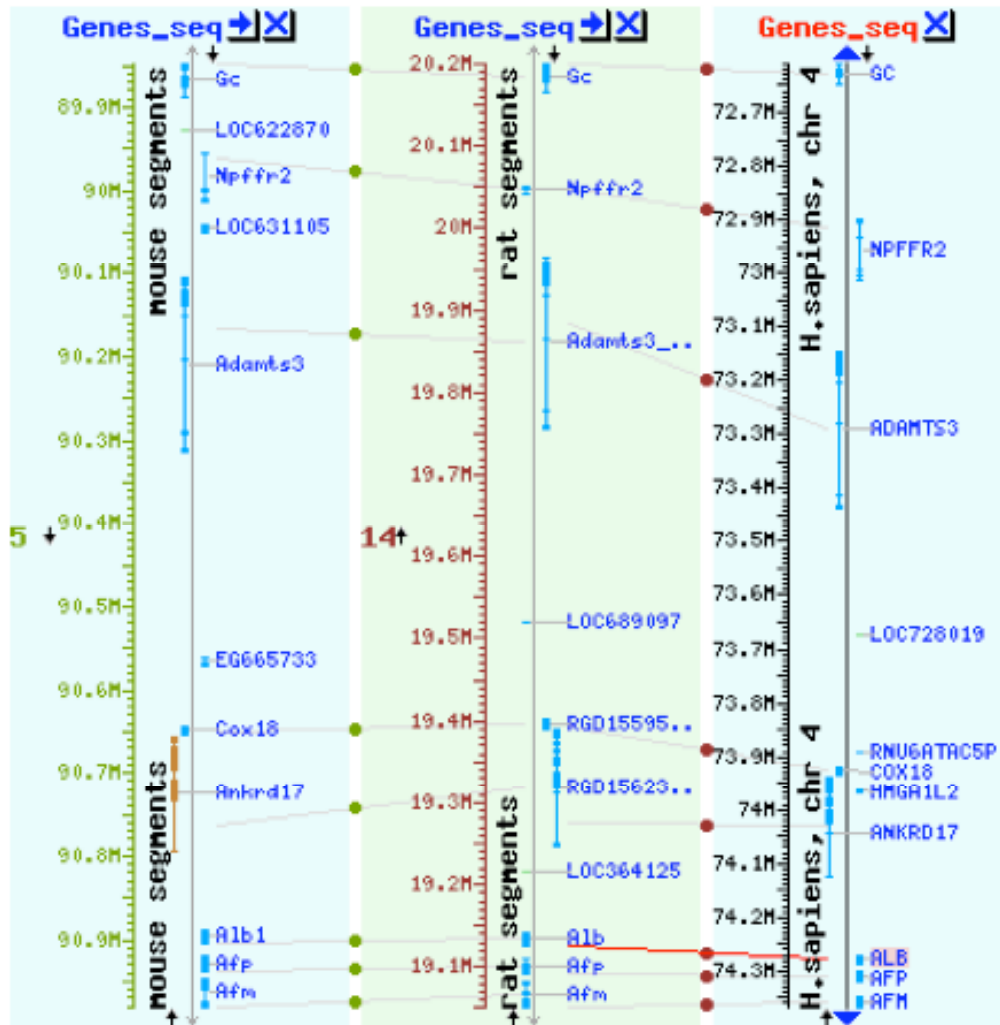
More Options:☒ Show Connections ☒ Verbose ModeCompress Map: Auto Compress if > pxPage Length: Thumbnail View: ☒ default (ideogram) ☐ master

OK

Apply

Close

This display shows gene-to-gene connections of the three genomes. Removing the UniGene map may make this easier to view. These connections are created through the HomoloGene database. Notice that the structure of the albumin gene family is conserved in these three mammalian genomes.



Genomic BLAST pages

Some of the higher genome BLAST pages are helpful because they allow the genomic context of the BLAST search to be displayed in the Map Viewer. We can use the human albumin RefSeq transcript to identify the homolog in the rat genome.

Follow the link from the BLAST home page (<http://www.ncbi.nlm.nih.gov/blast/>) to the rat genome BLAST page.

Type the accession number for the human albumin precursor, NM_000477, into the search box on the BLAST form.

BLAST Rat Sequences. - Mozilla Firefox

File Edit View History Bookmarks Tools Help

NCBI Home ► Genomic Biology ► Rat Genome Resources ► BLAST

Search Map Viewer Go Clear

BLAST
Overview
FAQs
News
Manual
References
Retrieve results
Genome Project

BLAST Rat Sequences.

☒ Enter an accession, gi, or a sequence in FASTA format:

☐ Or, choose a file to upload

Set subsequence: (optional)
From: To:

Database:
genome (all assemblies) 8014 sequences

Program:
megaBLAST: Compare highly related nucleotide sequences
cross-species megaBLAST: Compare nucleotide sequences for other species to this genome
BLASTN: Compare nucleotide sequences
BLASTP: Compare protein sequences
BLASTX: Compare a nucleotide sequence against a protein database
TBLASTN: Compare a protein sequence against a nucleotide database

0.01 default 100 100

Advanced options:

Run the search without changing the default settings.

This will use megablast against the assembly. This is faster but less sensitive than ordinary blastn when run in contiguous word-hit mode (word size =28, exact match required) as it is here.

Format your results.

Were you able to find the rat homolog?

Repeat the search. This time choose the cross-species megablast option.

You should have found some hits this time. The graphical overview shows that some parts of the human albumin transcript did not find any significant matches in the rat. Albumins are not highly conserved genes. Notice that the alignments shown in the output are some of the exons of the rat albumin gene. The exon matches are ordered by significance; the longest and best conserved exons are shown first. Another more interesting way to display these is by the position in the genome.

Display your results in the rat Map Viewer by linking through the linked RefSeq identifier to the contig on rat chromosome 14.

Notice that not all exons were found and that none of the other gene family members were identified. You can potentially identify the other members of the gene family in rat by searching with the human protein using the translation of the genome

Go back to the rat genome BLAST page and type the human RefSeq protein accession, NP_000468, in the search box.

Select the translating BLAST search, tblastn, from the program selection drop down menu.

Display these results in the Map Viewer as with the nucleotide results.

You may need to adjust the zoom level to see your results clearly. What albumin gene family members did you find? If you compare the corresponding region in the human genome and mouse genomes you may notice that the rat and the mouse have an additional member close to afamin.

Albumins are also present in other vertebrates. We can use the specialized genomic BLAST pages to try to find homologs in other organisms. Links to some of the genome specific BLAST pages are available from the BLAST home page (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>). For some genomes, it's necessary to follow the BLAST link from the Map Viewer homepage. This is the link on the BLAST page labeled "[list all genomic BLAST databases](#)".

BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

- | | | |
|---|--|--|
| <input type="checkbox"/> Human | <input type="checkbox"/> Oryza sativa | <input type="checkbox"/> Gallus gallus |
| <input type="checkbox"/> Mouse | <input type="checkbox"/> Bos taurus | <input type="checkbox"/> Pan troglodytes |
| <input type="checkbox"/> Rat | <input type="checkbox"/> Danio rerio | <input type="checkbox"/> Microbes |
| <input type="checkbox"/> Arabidopsis thaliana | <input type="checkbox"/> Drosophila melanogaster | <input type="checkbox"/> Apis mellifera |

This links to the Map Viewer Homepage (<http://www.ncbi.nlm.nih.gov/mapview/>).

Follow the link from the BLAST or Map Viewer home page to the chicken (*Gallus gallus*) genome BLAST page.

Type the accession number for the human albumin precursor, NP_000468, into the search box on the BLAST form.

Select the translating BLAST search, **tblastn**, from the program selection drop down menu.

Leave the database menus set on “genome” and click the BLAST button.

Format your results.

Were you able to find matches in the chicken genome? How many potential homologs did you find?

Repeat the search for the albumin family using the horse genome BLAST page. The horse genome BLAST page is linked through the “B” icon on the Map Viewer homepage.

Gene and RefSNP: Craig Venter’s APOE status

Certain isoforms (allelic variants) of the Apolipoprotein E gene are associated with increased risk for cardiovascular disease and late-onset Alzheimer disease. The common alleles are shown in the table below

Isoform	Position 112 (130)	Position 158 (176)
e3	Cysteine (C)	Arginine (R)
e4	Arginine (R)	Arginine (R)
e2	Cysteine (C)	Cysteine (C)

The positions in parenthesis include the leader peptide, and will be the positions in the unprocessed protein. The epsilon 4 isoform seems to increase risk relative to the epsilon 3 allele, while the rarer epsilon 2 allele seems to be protective

Search Gene for APOE and retrieve the human Gene record. Follow the link to Gene View in dbSNP to see the reference genome status at the critical positions.

Which allele is represented in the reference assembly?

Follow the link to the RefSNP at position 130. Look at the integrated maps section to see the status in J. Craig Venter’s genome (HuRef). You can verify this in the Graphic Sequence viewer by loading NW_001838496, searching for APOE, and zooming in on the last exon.

Using NCBI BLAST

Identifying sequences

Michael Crichton's fantasy about cloning dinosaurs, *Jurassic Park*, contains a putative dinosaur DNA sequence. Use basic nucleotide BLAST against the nucleotide database, nr, to identify the real source of the following sequence from the novel. You can retrieve the sequence from the NCBI ftp site:

<ftp.ncbi.nih.gov/pub/FieldGuide/jurassic.txt>

Select, copy and paste the sequence into the BLAST form window and run the search against the nr(nt) database. Use the default Megablast algorithm.

What is the sequence that Michael Crichton used?

This search is an example of the most common use of nucleotide-nucleotide BLAST: sequence identification, establishing whether an exact match for a sequence is already present in the database.

Mark Boguski, who was at the NCBI at the time, noticed this obvious contaminant and supplied Crichton with a better sequence for the sequel, *The Lost World*. You can also retrieve this sequence from the NCBI ftp site:

<ftp.ncbi.nih.gov/pub/FieldGuide/lostworld.txt>

Select, copy and paste the sequence into the BLAST form window and run the search.

Identify the most likely source of this sequence using nucleotide-nucleotide BLAST.

Mark imbedded his name in the sequence he provided. To see Mark's name, use the translating BLAST (blastx) page with the sequence. (Look for MARK WAS HERE NIH).

The most important use of the translating BLAST services is to look for similar proteins (identify potential homologs) in other species.

Short Nucleotide Sequences and Algorithm Parameters

A frequent use of nucleotide-nucleotide BLAST is to check the specificity oligonucleotides for hybridization or PCR. The goal most people have when doing this is to make sure that the primer will give a unique product from the target genome or cDNA population. Because BLAST is local

and searches both strands, one can simply concatenate a pair of +/- strand primers and use them in a single search. You can try the traditional method below with this set of primers then use the new PrimerBLAST tool, linked under "Specialized BLAST" to check the primers

Combine the following pair of candidate PCR primers in a nucleotide-nucleotide search against the nr(nt) database. Be sure to choose blastn (Somewhat similar sequences) as the BLAST program under "Program selection."

F12 GTCAAGTGGCAACTCCGTCAG

R8 TTGAGAGATGGATTGTTGCTC

To prevent false matches that overlap the forward and reverse primer sequences, type ten or more "n's" between the sequences when using them as a query.

GTCAAGTGGCAACTCCGTCAGnnnnnnnnnnTTGAGAGATGGATTGTTGCTC

Retrieve the results and identify the gene amplified by these primers.

What is the predicted size of the product that would be amplified by PCR from cDNA (RT-PCR)? How could you distinguish the products amplified from genomic DNA versus cDNA?

You can also try these primers against the human genomic plus transcript database to get a clearer view of the product predicted from genomic DNA in the Map Viewer.

Now try these modified primers in standard nucleotide-nucleotide BLAST. There is one mismatch in each near the middle.

F12_mod GTCAAGTGGCgACTCCGTCAG

R8_mod TTGAGAGATGtATTGTTGCTC

GTCAAGTGGCgACTCCGTCAGnnnnnnnnnnTTGAGAGATGtATTGTTGCTC

Notice that the previous hits are completely missing. This is because the default word size setting requires an exact match of 11 before extensions can occur. A mismatch in the middle of a 21-mer will prevent any initial word hits. There is an automatic adjustment for short sequences that will allow these hits with mismatches to be found. However the sequence with the linking "n's" is too long to trigger the adjustment.

Run the search again with the forward and reverse primers as separate sequences. Copy and paste the following FASTA formatted primers in the search box.

```
>F12_mod
GTCAAGTGGCgACTCCGTCAG
>R8_mod
TTGAGAGATGtATTGTTGCTC
```

Your results should now display a message that your search parameters were adjusted to search for a short input sequence, and you should see results for both primers. Notice that although there are now hits, the original hits are still missing. This is because the expect value of the mismatch hits is above 10.

You can manually adjust search parameters to short sequence setting through the “Algorithm parameters” section of the nucleotide BLAST form. After adjusting these, the search with the concatenated mismatched primer will work.

Go back to the BLAST form. Click on the reset page link at the top to restore the default settings. Then select blastn under “Program Selection” and expand the “Algorithm parameters” section of the form. Make the following changes.

- **Uncheck the box next to “Automatically adjust parameters for short input sequences.”**
- **Increase the expect threshold to 100.**
- **Set the Word size to 7**
- **Set the Match Mismatch Scores to 1, -3**
- **Uncheck any Filter options**

Now run the search again with the concatenated mismatch primers.

GTCAAGTGGCgACTCCGTCAGnnnnnnnnnTTGAGAGATGtATTGTTGCTC

Algorithm parameters

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: ☒ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 11

Scoring Parameters

Match/Mismatch Scores: 2,-3

Gap Costs: Existence: 5 Extension: 2

Filters and Masking

Filter: ☒ Low complexity regions
☒ Species-specific repeats for: Human

Mask: ☒ Mask for lookup table only
☐ Mask lower case letters

Algorithm parameters Note: Parameter values that differ from the default are highlighted in yellow

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: ☐ Automatically adjust parameters for short input sequences

Expect threshold: 100

Word size: 7

Scoring Parameters

Match/Mismatch Scores: 1:3

Gap Costs: Existence: 5 Extension: 2

Filters and Masking

Filter: ☒ Low complexity regions
☐ Species-specific repeats for: Human

Mask: ☒ Mask for lookup table only
☐ Mask lower case letters

Do you find the original hits now?

Protein-protein BLAST and Short Peptides: ELVIS lives

As the database grows, so does the number of chance occurrences of amino acid motifs that spell out words or people's names in single-letter amino acid codes. One such name motif is ELVIS. In this example we will count the number of occurrences of ELVIS in the default protein database. The automatic adjustment of search parameters will allow us to find matches with this short peptide

Type **ELVIS** in the search box on the **blastp** form.

Expand the **Algorithm parameters** section and adjust the number of **Max target sequences** to **1000** or more to include all Elvises.

Run the search.

What is the expect value for an exact match to ELVIS? The number of Elvises increases in a linear fashion with the size of the database in accordance with the random behavior of protein sequences.

Click on the **“Edit and Resubmit”** link at the top of the **BLAST** form. Examine the **Algorithm parameters** section to see how the settings were adjusted to search with this short peptide.

PSI-BLAST and Conserved Domains

The histidine kinase-like ATPase domain (HATPase_c) is present in a wide variety of proteins with quite different functions. These include bacterial sensor histidine kinases, DNA mismatch repair proteins, topoisomerases, DNA gyrases and 90 KDa heat shock protein homologs. We can use PSI-BLAST to demonstrate the similarity among these proteins that is not apparent with ordinary BLAST.

Use the human DNA mismatch repair protein MLH1 (NP_000240) in an ordinary blastp search and examine the conserved domain results to verify the presence of the HATPase_c domain.

From the results of the above search, click the “Edit and Resubmit” link and make the following changes to prepare to run a PSI-BLAST search with just the region of MLH1 that corresponds to the HATPase_c domain

- **Set the query subrange in the boxes on the right hand side of the form. Use 32 as the “From” coordinate and 122 as the “To” coordinate.**
- **Change the database to “swissprot.”**
- **Change the “Program Selection to PSI-BLAST.”**
- **Expand the “Algorithm parameters” section and set the “Max target sequences” to 5000.**

Now click the BLAST button to run the first iteration of PSI-BLAST and examine the results.

The results are just the blastp results that are formatted for PSI-BLAST. Notice that the descriptions section of the results is divided into two sections. The upper section contains the sequence with alignments that will be used to generate the position specific score matrix in the next iteration of PSI-BLAST. These sequence alignments have e-values less than 0.005. This cut-off is empirically determined to give good results in PSI-BLAST searches. All of the proteins above this threshold in the first iteration are DNA mismatch repair proteins PMS, MutL and HexB homologs. Just below the PSI-BLAST threshold with e-values ranging from 0.008 to 6.0 are several bacterial signaling histidine kinases. Some of these have marginally significant e-values in ordinary BLAST but many are not distinguishable from chance matches.

Now click the “Run PSI-BLAST iteration 2” button to run the second iteration of PSI-BLAST and examine the results.

There are now new proteins less than the 0.005 threshold. Notice that these are now marked with a “New” graphic while the proteins found in the previous iteration are marked with a green ball. Many of the new proteins are topoisomerases or DNA gyrases. There are also many more gyrases and topoisomerases just above the 0.005 threshold.

Retrieve a few of the new proteins in Entrez by clicking on the linked identifier and verify that they contain the HATPase_c domain.

Click the “Run PSI-BLAST iteration 3” button to run the third iteration of PSI-BLAST and examine the results.

Again there are new proteins, not only more gyrases and topoisomerases, but also signaling histidine kinases and HSP90 chaperonins.

Continue to run PSI-BLAST iterations until you have collected some plant phytochrome and ethylene receptor proteins below the 0.005 threshold.

These are plant signaling proteins. As these results show, plant ethylene receptors and phytochromes are related by sequence similarity to the two component sensor kinase system of bacteria.

Demonstrate the similarity between the HATPase_c domain of the *E. coli* sensor protein PhoR (PHOR_ECOLI, P08400) and plant ethylene receptors by performing a first iteration PSI search against swissprot. Use a query subrange on PhoR of 318 to 421.

Now, continue to run PSI-BLAST iterations until the plant phytochromes appear.

The number of iterations should be fewer than when using the MLH1 protein as a query.

Retrieve the protein record for an ethylene receptor (ETR1_LYCES, ETR1_ARATH) and a phytochrome (PHYA_ARATH, PHY_PICA) by following the link to Entrez. Compare their domain structures by following the links to the pre-computed Conserved Domains results.

What three domains do they have in common?

Retrieve a protein record for one of the bacterial sensor proteins (PHOR_ECOLI) and examine its domain structure.

Notice that the plant proteins and the bacterial protein all contain the histidine kinase domain (HATPase_c) and the HisKA (phosphoacceptor) domain. In the classic two component bacterial system, the HisKA domain is phosphorylated on a conserved histidine residue by the HATPase_c domain in response to an external signal. This phosphate is then transferred from the HisKA domain of the sensor protein to a conserved receiver domain on a separate response regulator protein. In the case of PhoR the response regulator is PhoB.

Retrieve the *E. coli* PhoB protein (PHOB_ECOLI, P0AFJ5) and examine its domain structure as before.

Notice the presence of the receiver domain (REC) and a DNA binding effector domain (trans_reg_C) in PhoB. In the plant ethylene receptors examined previously there is a receiver domain is on the receptor itself, but the effector domain present in PhoB is lacking. The plant ethylene receptors apparently mediate their effects through the MAP-kinase pathway. Unlike the ethylene receptors, the phytochromes function as serine/threonine kinases but also appear to share an ancestry with bacterial histidine kinases.

Translating BLAST searches, mining polymorphisms

The prion protein is found in high concentrations in the brains of humans and other mammals. In certain degenerative neurological diseases, prion proteins aggregate into polymers. Several of these prion diseases seem to be transmissible. Perhaps the most remarkable aspect of these is that the infectious agent appears to be an aberrant form of the prion protein itself. Bovine spongiform encephalopathy (BSE) is one of the transmissible prion diseases that has received much recent notoriety. There are a number of polymorphisms that have been identified in the prion proteins for several mammals, notably human, mouse, and sheep. Some of these are associated with inherited prion diseases and some with susceptibility to transmissible forms.

Retrieve the SWISS-PROT record for the human prion protein (PRIO_HUMAN, P04156) and look at the FEATURE table to see the various polymorphisms.

Notice the methionine / valine polymorphism at position 129. The amino acid at this position affects the particular disease phenotype when another disease causing mutation is present. People who are heterozygous at this position appear to be more resistant to *kuru*, one of the transmissible encephalopathies. There is population genetic evidence that there may have been balancing selection for heterozygotes at this position during human evolution. The EST data for human represents a large number of individuals and can be used as a resource for identifying nucleotide polymorphisms. In this case, we can investigate the prevalence of the two alleles at position 129 of the prion protein in the EST data for human. We will use one of the formatting options to make the different alleles easier to identify.

Set up and run this search by following these steps:

- **From the BLAST homepage, link to the tblastn form “Search translated nucleotide database using a protein query.”**
- **Type the prion protein accession number, P04156, in the search text area.**
- **Use the “Query subrange” boxes to use only residues 100 to 160.**
- **Choose the “Expressed sequence tags (est)” database.**
- **Type “human” in the Organism limit box and choose human (taxid:9606) from the resulting list to limit to human sequences.**
- **Open the Algorithm parameters section and set the Max target sequences to 1000**
- **Turn off the “Low complexity” filter option.**
- **Click the BLAST button to run the search.**
- **Immediately click the “Formatting options” link at the top of the intermediate page.**
- **Set the alignment view to “Query-anchored with dots for identities.” This is a stacked pairwise alignment format that makes it easy to see changes relative to the query sequence in all the database hits at once.**

Click the “View report” button to display the results.

Look at the alignments to see how the query-anchored format helps to investigate changes in sequences. Find position 129 in the query. Which amino acid is most prevalent at position 129?

WGS and Trace Archive Data in Entrez and BLAST

Verify that nearly all of the rabbit DNA records in the NCBI database are whole genome shotgun. You can retrieve all nucleotide rabbit sequences by using the Limits tab and setting the field restriction in the pull-down list to organism. You can further limit to genomic DNA through the “molecule” pull-down list.

How many records are there?

Follow the link to the “CoreNucleotide” results before continuing. Now restrict to whole genome shotgun records by adding the following query term to your search.

wgs[Properties]

The overall search performed now is

rabbit [Organism] AND biomol_genomic[Properties] AND wgs[Properties]

The first record is the master record for the project that gathers all of the contigs. You can get only this record by adding wgs_master[Properties] to the search.

Retrieve the first contig record in your list and verify that it is unannotated –no genes or other features.

Using BLAST, Spidey and Splign to annotate wgs

You can find the genomic sequences corresponding to a rabbit (*Oryctolagus cuniculus*) mRNA sequence by using BLAST to search the wgs database. A sequence that demonstrates this is the rabbit apolipoprotein A-1 mRNA (NM_001101687).

- **From the BLAST homepage select the blastn page**
- **Type NM_001101687 in the “Search box” and select wgs as the database.**
- **Use the Organism limit feature to limit to rabbit (taxid:9986)**
- **Expand the Algorithm parameters section and set the e-value threshold to 1e-12.**
- **Run the search and re-format your results using the “CDS feature” option and “Pairwise with identities” Alignment view option.**

Your search should hit one wgs contig (AAGW01335306). How many exons did you identify in each?

Use the sort by “Query start position” to put the exons in the genomic order on AAGW01335306.

This is a rather primitive gene model because it does not constrain the alignment breaks to splice junctions.

Use the same mRNA and genomic sequences as above, make gene models using the spliced alignment tools Spidey and Splign and compare them to the BLAST results.


The spliced alignment tools place two of the exon-intron boundaries at slightly different points than BLAST alignments.

Trace Archive

Some sequences are only available through the NCBI trace archive.

<http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?>

These data can be retrieved by species code or trace number. The most important way to search these is through the Trace archive megablast pages. Both standard (contiguous) megablast and cross-species megablast are available. These are linked through the BLAST tab on the main trace archive page or through the BLAST homepage in the Specialized BLAST section.

 **Trace Archive**

[Main](#)
[Obtaining Data](#)
[Statistics](#)
[Tracking](#)
[Documentation](#)
[Trace Assembly](#)
[SRA](#)
[Trace Home](#)
[Trace BLAST](#)

[News/Events](#)
[FAQ](#)
[Acknowledgment](#)
[FTP](#)

Enter a *query string* ([use Query Builder](#)) or *TI number*

alt ☐

Last week Top 10 Arrivals (10/25/2009 - 10/31/2009)

Organism	Count
HOMO SAPIENS	1,819,859
HUMAN GUT METAGENOME	595,201
SUS SCROFA	113,473
HUMAN METAGENOME	55,242
CHRYSEMYS PICTA	43,916
BODO SALTANS	16,060
CAVIA PORCELLUS	13,048
OCHOTONA PRINCEPS	10,486
MOUSE GUT METAGENOME	8,739
CALLITHRIX JACCHUS	8,453

We can use the cross-species page to find an HSP70 gene homolog in the sea lamprey (*Petromyzon marinus*) traces

- Go to the **BLAST homepage** and choose the trace archive search from the **Specialized BLAST** section.
- Enter the accession number for the human HSP70 1A mRNA Reference Sequence (NM_005345) in the search box on the BLAST form.
- Choose **Petromyzon marinus-WGS** as the database and set **blastn** as the program.
- Click the **BLAST** button to run the search.

Because HSP70 is well conserved it is easy to find homologs in the sea lamprey at the nucleotide level using the human sequence. Many less well-conserved genes may only be identified at the protein level. Unfortunately the large size of the trace databases makes translating searches impractical.

New BLAST Displays

TreeView

The treeview display in BLAST will not always produce reasonable phylogenetic species trees or gene trees because the alignments are not multiple sequence alignments and don't necessarily

include all residues. Nevertheless searches with complete mitochondrial genomes often reproduce accepted phylogenetic groupings.

From the BLAST homepage, choose the blastn page. . Select the RefSeq genomic database from the database pull-down list and put the accession for the wolf mitochondrial genome (NC_008092) in the “Search” box as a query.

The Refseq genomic database contains chromosome (NC_) RefSeqs including plastid genomes, mitochondrial genomes and chromosomes for prokaryotic genomes.

Use the following Entrez limit to restrict to the mammalian order carnivora (dogs, cats, seals, hyenas, weasels etc.).

carnivores[organism] NOT gene in genomic[properties]

This last term, “NOT gene in genomic[properties]”, eliminates hits to mitochondrial insertion sequences present in the dog genome.

Run the search. Click on the “Distance Tree of Results” link under the BLAST graphic to display the tree. Compare the groupings to the classification of the carnivores in the NCBI Taxonomy database.

The family groupings correspond to those in the tree. However, many families of carnivores are not represented because the mitochondrial genomic sequences are not available yet.

New View of Results and Genome and Transcript Databases

The new human genome and transcript database provides direct access to the human genome through the main BLAST page. The new view options provide a more organized and sortable presentation of the results.

Use nucleotide-nucleotide BLAST (blastn) to search the human genome plus transcript database with the human alcohol lactate dehydrogenase B (LDHB) transcript (NM_002300).

Use the new sorting options and summary statistics to identify the functional multi-exon gene by sorting using the “Total Score” column.

On which chromosome is the functional gene? How many exons does it have?

Use the “Sort alignments” feature to sort by “Query start position” to get the exons in genomic order.

On which chromosome is the longest retrocopy pseudogene?

Follow the linked identifiers to the human Map Viewer to display the hits for both the functional gene and the retrocopy pseudo gene.